

Strategy for Framework 6

The European Commission is embarking on plans for its next Framework programme of research and development. Linking it to a broader strategy is advisable, provided political goodwill is preserved.

With half-time just passed, the performance of the European Commission's fifth Framework programme of research has been weighed — and found wanting. Scientists, research lobbyists and review panels all complain about its complicated rules and application procedures. The commission itself now appreciates the depth of the concern over the programme, and that it tends to discourage scientists with its need to demonstrate relevance to socio-economic goals (see *Nature* **398**, 1; 1999).

What is to be done? Don't start from here, is one response. Over the next few weeks, the commission will be developing the outlines of the sixth Framework programme (FP6), due to start in 2003. Research commissioner Philippe Busquin is intent on basing FP6 on new foundations by interweaving it with the concept of a single 'European research area' (see *Nature* **405**, 873; 2000).

Under FP6, the commission plans to play a more strategic role. Alongside simplifying the rules, concentrating on fewer priorities and decentralizing project funding, more emphasis will be placed on creating equal conditions and opportunities for scientists throughout the European Union. The commitment to improving Europe's human research potential and technological innovation will continue. Administrative issues, such as planning and co-financing joint research infrastructures, are intended to take precedence

over direct funding of small, scattered research projects.

But these visions will only materialize if Busquin manages to bring the member states on board. The funding priorities under FP5 — the life sciences, information technology, energy and the environment — deserve further multinational support. If the commission draws back on project funding, the only means of replacing it will be from the member states' budgets. That is why Busquin tirelessly emphasizes the importance of networking and the joint execution of national funding agencies' programmes. But although he has met with positive responses from member states so far, the FP6 plans may put a stop to all that. For legal, financial and perhaps also political reasons, many national funding agencies are not ready to open their programmes to researchers Europe-wide. Moreover, there is insufficient incentive to integrate national funding efforts, despite expressions of goodwill.

A step-by-step approach is essential, therefore. The single European research area is a noble goal, and FP6 is a major opportunity in that direction. But it would be unwise to put too much hope in the goodwill of the member states. Under an improved and streamlined process, project funding should remain a key task for the commission — at least until independent pan-European science agencies are established. Regrettably, all the signs are that that is more than a Framework programme away. ■

Pigs, society and opacity

International agencies need to learn the lessons of the past about ill-advised secrecy.

The ethics and safety of xenotransplantation will be debated by experts, representatives of governments and international organizations at a three-day meeting in Paris this week. One likely outcome is a joint position paper and recommendation by the Organisation for Economic Co-operation and Development (OECD) and the World Health Organization (WHO). Pre-meeting information points out that, although clinical trials of xenotransplantation are already happening (notably, of fetal pig neural cells and pig liver cells), the issues, and in particular the risks of infectious disease spreading to humans, "have yet to be fully addressed internationally". The issue is contentious, with huge private investment fuelling a rush to clinical trials while others want a moratorium on trials.

But journalists will be denied access to the meeting, and must make do with a press conference at the end. One lesson from the BSE crisis and the controversy over genetically modified organisms is that transparency is the key to obtaining public confidence in the process of drafting recommendations on areas where risks exist alongside scientific uncertainty. Like it or not, the media (not all of which are mischievous or incompetent) remain the principal channel for transmitting information to an increasingly concerned public, and for analysis of the complex issues involved.

Moreover, in the public interest, journalists must be aware, at first hand, of the differences in the views aired. International organizations have often published the proceedings of such meetings only after considerable delays, and when the texts have been stripped of

the more controversial issues as national governments have been given the right to censor them.

An opportunity to remedy this has been squandered by the organizers of the Paris meeting. Media participation at such meetings should be encouraged, and organizers should invest in providing comprehensive background material in comprehensible language to help journalists get up to speed. Openness carries risks: complex issues may be misunderstood or misrepresented. But in the long run it is preferable to closed debate.

A request by a *Nature* journalist to attend the meeting met with the explanation that it was originally to have been open to the media, but that this was vetoed by WHO and governments, some of whom argued — incorrectly — that "journalists weren't interested in spending hours and hours in meetings".

OECD staff appear dismayed by their partners' decision, and admit that this raises an issue: the need for international organizations to develop clear and mutually acceptable policies *vis-à-vis* media participation. Indeed, and not a moment too soon. Meetings of committees that offer advice to the US government are required to be open under the Federal Advisory Committee Act. In fact, there is no reason that meetings of international agencies should not be broadcast live on the web for all to judge. Both OECD and WHO have the technical capacity. What is stopping them? Inertia, perhaps. But it is surely an obsolete notion that risk is best handled in the closed corridors of selected 'experts' and government agencies, far from the public eye. ■





Legal headache

Tug-of-war over rights to skeleton of Native American
p548



Coming together

New lease of life for Japanese fusion experiment
p549



Warming up

More samples ready as Stone Age iceman defrosted
p550



Tough decision

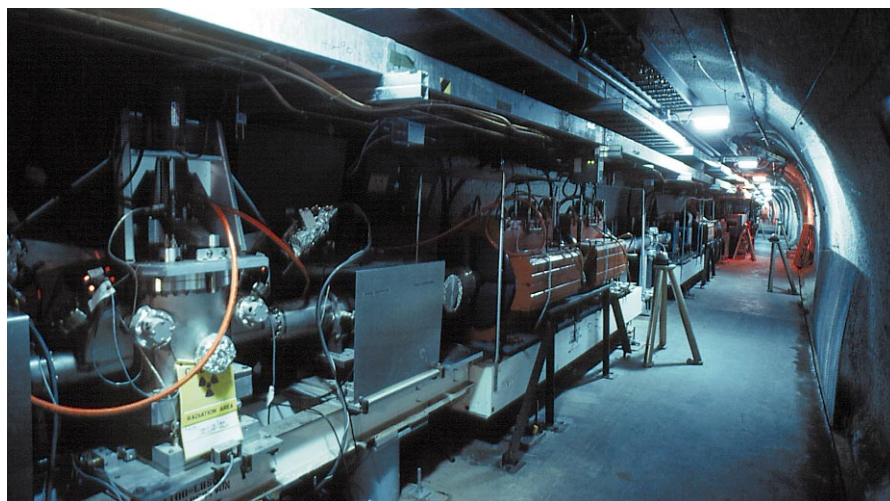
US finally gives abortion pill the go-ahead
p553

US considers moves to relieve morale crisis at energy labs

Colin Macilwain, Washington

Ever since March 1999, when allegations of Chinese spying at the Los Alamos National Laboratory in New Mexico burst into public view, draconian security measures have been imposed at Los Alamos and other laboratories run by the US Department of Energy (DoE). But following the release of Wen Ho Lee, the Taiwan-born scientist accused of passing on nuclear secrets, on 13 September, scientists are cautiously optimistic that the measures — which have caused a crisis of morale at the labs (see *Nature* 407, 447–448; 2000) — will begin to be rolled back.

Last week, several developments suggested that the worst of the storm may be passing and that Washington is preparing to take steps to repair the damage caused by the espionage investigation. On 25 September, a report prepared by former Republican senator Howard Baker and former Democrat representative Lee Hamilton — two highly respected figures in Washington — denounced the impact of recent Federal Bureau of Investigation inquiries into the temporary disappearance of two computer hard drives at Los Alamos in May. “The most worrisome known consequence of the hard-



Energy stall: the Stanford Linear Accelerator Center does no classified work, but is still restricted.

drive incident is the devastating effect on the morale and productivity of the laboratory,” their report said.

Two days later, at a National Academy of Sciences (NAS) meeting on scientific communication and national security, Neal Lane, science adviser to President Bill Clinton, acknowledged that new security controls at

the laboratories “have overshoot the mark”. Lane said that scientists there “have been subjected to sensational allegations by the press and by Congress”, leading to “a siege mentality at the laboratories”.

The next day, negotiators from the two houses of Congress agreed a budget bill for the DoE that would restore the sum of money that the weapons labs’ directors can spend on science at their own discretion from 4% to 6% of their budgets. The reduction of this fund a year ago had curtailed some of the basic, non-classified research that attracts top scientific talent to the laboratories (see *Nature* 402, 449; 1999).

But the increased bureaucracy resulting from the security clampdown is also causing problems at DoE labs that do not do classified research. “There are two distinct problems,” explains Wolfgang Panofsky, a leading physicist and veteran of the Los Alamos Manhattan Project, now at the Stanford Linear Accelerator Center (SLAC) in California. “The weapons laboratories are demoralized, and the rest of the DoE laboratories have all this paperwork and regulation to deal with.”

Energy department officials now say that they intend to work with the National Nuclear Security Administration (NNSA), the semi-autonomous agency established to

Novartis axes UK transplant centre

Jessa Netting

The Swiss drug company Novartis is set to sidestep mounting controversy over its xenotransplantation research in Britain by shutting its UK subsidiary Imutran. It is collaborating with the US-based company BioTransplant to set up a company that will conduct the research in Massachusetts.

Imutran researchers have transplanted pig organs into monkeys, with a view to developing the technique for use in humans. But some scientists are concerned about the potential for retroviruses — inactive and harmless in the pig genome — to become infectious in humans. In addition, animal rights activists have campaigned against Imutran’s research.

The Campaign for Responsible

Transplantation, a pressure group critical of xenotransplantation, branded the move by Novartis as an attempt to duck the negative public sentiment in Europe and to take advantage of a more relaxed regulatory climate in the United States.

But Paul Herrling, head of global research at Novartis, dismisses the idea that the plan was a response to recent bad publicity. He does concede, however, that the United States already has defined regulatory criteria that would allow human trials.

Novartis will invest \$30 million over three years to take a two-thirds stake in the new company, with BioTransplant owning the other third. It says all relevant Imutran operations will move to the new company. ■

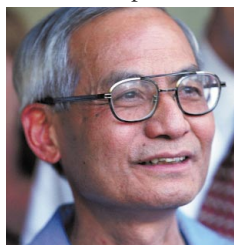
http://www.crt-online.org

run the weapons laboratories, to make a new start in their efforts to improve both security and morale. "We realize that the way things are now is not going to work," says Marshall Combs, an adviser to energy secretary Bill Richardson. "This is going to be a new time. We are going to pull this off."

But researchers and managers at the DoE labs warn that rebuilding morale will be difficult. Los Alamos suffered another blow last week, with reports that senior staff there, including the director John Browne, will be disciplined by the University of California for their role in the hard-drive incident. At the NAS meeting, ample evidence was presented of the toll the crisis is taking at its weapons labs as well as at the DoE's huge network of non-weapons labs.

Jonathan Dorfman, director of SLAC, which does no classified work at all, says the agency nonetheless faces "an unrelenting affront to the open research environment" characterized by "ill-conceived, one-size-fits-all directives". In June 1999, he says, there were 11 security directives pertinent to SLAC; now there are 31, with 21 more in draft. According to Ned Sauthoff of the Princeton Plasma Physics Laboratory, one draft asks researchers to get permission before making an international phone call.

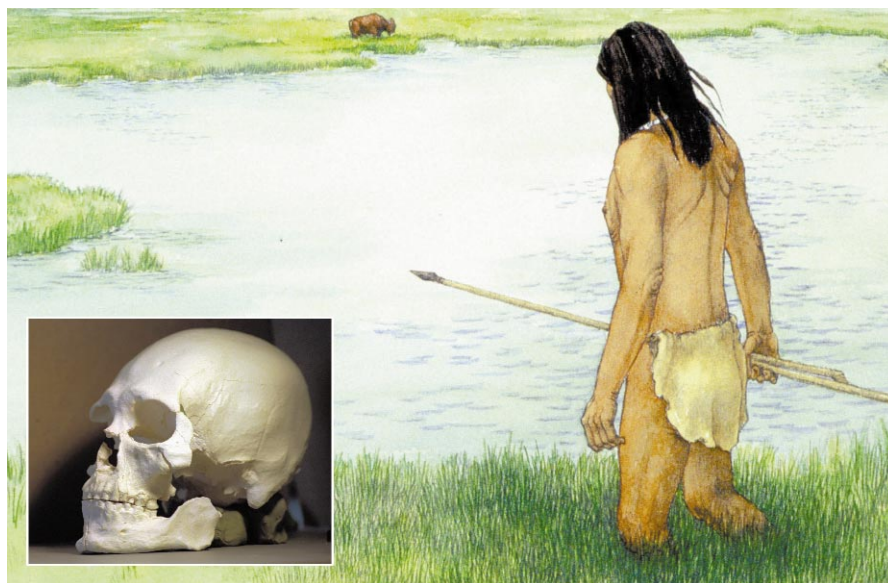
Dorfman wants laboratories such as SLAC exempted from the rules. But



Lee: release from prison has prompted optimism.

the case for exemption is complicated by the intermediate status of predominantly civilian laboratories such as Brookhaven in New York and Argonne in Illinois, which perform small amounts of classified work.

The best hope for a new beginning at the laboratory complex is perhaps the congressionally mandated NNSA, which has assumed responsibility for all of their nuclear weapons-related work. Will Harper, the Princeton University physicist and former assistant energy secretary, who organized last week's NAS meeting, predicts that the NNSA will prove effective — once it establishes autonomy on a par with that enjoyed, for example, by the National Institutes of Health within the health department. John Gordon — a former physics researcher and later a general in the Air Force — was appointed this summer as the NNSA's first administrator. Whereas other senior posts in the government will rotate after next month's elections, Gordon has been told by both political parties that he will serve for at least three years, ensuring continuity. ■



Culture clash: an artist's impression of Kennewick man, whose remains are caught up in a legal battle.

Researchers fight for access to Native American skeleton

Rex Dalton

Defiant researchers are taking the US government to court over its plan to return the much-debated 'Kennewick man' skeleton to Native American tribes for reburial.

Nearly four years ago, eight researchers sued the federal government to secure the right to study the ancient specimen, which was found in 1996 on federal land by the banks of the Columbia River near Kennewick, Washington.

But last week, after a procedural delay in the lawsuit, the US Department of the Interior ruled that the skeleton, which is 80% complete, should be turned over to five Native American tribes for reburial. The decision was based on the Native American Graves Protection and Repatriation Act, which provides for the reburial of historic human remains.

The scientists now claim that the government is misinterpreting the law and blocking important research on one of the oldest and most intriguing specimens from the early peopling of the Americas, where only a dozen skulls over 8,000 years old have ever been found.

Partial analysis of the specimen indicates that it may be the remains of a man from Asia. The researchers want the legal right to examine the skeleton, and their lawyers will seek trial of the case at the US district court in Portland, Oregon.

Government lawyers declined to discuss the case, referring enquiries to the office of the interior secretary, Bruce Babbitt, where the decision was made. His spokeswoman, Stephanie Hanna, says Babbitt "feels all

relevant evidence has been gathered" and that it shows "a cultural affiliation" to the tribes that desire reburial.

But the government's decision was branded as "ludicrous" by the archaeologist Robson Bonnichsen, director of the Center for the Study of the First Americans at Oregon State University and a plaintiff in the lawsuit. Bonnichsen wants access to the skeleton to study its bone mineralization.

The skeleton's age — estimated by radiocarbon dating at more than 9,300 years — means there is no way to link it to the specified Native American tribes.

For the Native American tribes in Washington and Oregon, however, the reburial of Kennewick man has become a symbol of control of historic human remains and of their legal rights. Armand Minthorn, board member of the Confederated Tribes of the Umatilla Indian Reservation, called the scientists' research plans "blatant desecration of sacred human remains".

Douglas Owsley of the Smithsonian Institution's National Museum of Natural History, says the lawsuit "is not against Native Americans". Owsley, who is suing as an individual, and biological anthropologist Richard Jantz of the University of Tennessee at Knoxville, another party to the suit, want to take 65 cranial measurements, comparing the resulting three-dimensional picture of the skull with a computerized database of 4,500 previously collected specimens.

The possibility of an amicable agreement to allow both study and reburial appears remote, with both sides accusing the other of unwillingness to negotiate. ■

Reactor refit ignites debate on Japan's fusion strategy

Robert Triendl, Tokyo

The Japan Atomic Energy Research Institute (JAERI) plans to overhaul its JT-60 tokamak fusion reactor at a cost of ¥35 billion (US\$325 million). The task will take between five and ten years.

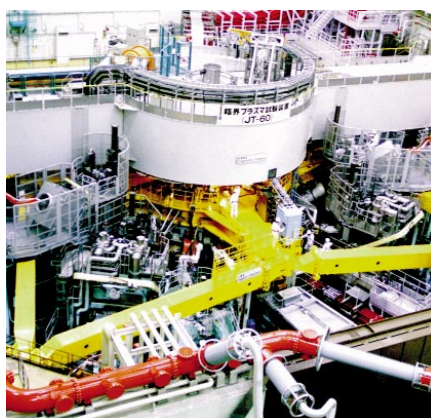
The reactor will be shut down in autumn of next year so that its magnetic coils can be replaced with superconducting ones. According to Masato Nakamura, director of the fusion development office at the Science and Technology Agency, of which JAERI is part, the goal is to increase the reactor's normal operation time from 15 to 100 seconds.

But what looks to be a boost for fusion might actually indicate that budgets for such research are growing tight in Japan, some observers say. The tokamak's closure will save a large part of its ¥10 billion annual operating costs — so the reconstruction could save the government money if it takes long enough. A completion date for the project has not been announced. And the Science and Technology Agency's budget request for 2001 includes only ¥2.3 billion for the reconstruction.

Fusion researchers are far from enthusiastic at the prospect of the reactor being shut down for up to ten years. "In terms of technology, replacing the coils should take three or four years and we hope this is what happens," says Shinzaburo Matsuda, director of JAERI's Naka Fusion Research Establishment.

Nakamura defends the plan, arguing that reconstructing the reactor will make it more economical and stable.

Both Nakamura and Matsuda insist that the plan will not affect Japan's commitment to the proposed International Thermonuclear Experimental Reactor (ITER) project.



Off line: refurbishment of Japan's JT-60 fusion reactor could take up to ten years.

"We believe that siting ITER in Japan is essential for the project's success," says Nakamura.

According to Nakamura, savings from the suspended operation of JT-60 will make it easier for Japan to increase its contribution to the ITER project.

Matsuda adds that, in its long term plan for fusion research, the Japanese Atomic Energy Commission defined an experimental reactor like ITER as the "core reactor" while designating "other tokamak devices" for advanced research to support the core machine during both the experimental and demonstration phases.

But Steve Dean of Fusion Power Associates, which lobbies for fusion research in Washington, expressed surprise at the plan. Although Japanese officials had said they could do both, outsiders had assumed that Japan would decide either to refurbish JT-60 or to proceed with ITER, he said. ■

Genomics initiative to decipher 10,000 protein structures

Paul Smaglik, Washington

In a bid to crack protein structures, the US National Institute of General Medical Sciences (NIGMS) last week launched the largest project to date to find the best methods for unravelling the molecules' three-dimensional make-up.

As part of its structural genomics initiative, the institute will provide \$150 million over five years to seven regionally based research groups. The NIGMS aims to solve the three-dimensional structure of 10,000 proteins — each representing a protein family — over the next decade.

Structural genomics compares how proteins of known structure share functions or gene sequences with unsolved proteins. It then uses computation to predict structures for the unsolved proteins.

But to assemble a full protein library, structural biology's technology and methodology need to improve. To that end, the seven groups will use robotics to automate protein characterization. Some will also seek to expedite or eliminate steps in sample preparation — the biggest bottleneck in large-scale characterization.

Each of the groups is using a different rationale for choosing which proteins to characterize. They will study proteins from different organisms, ranging from yeast to humans. Some will use X-ray crystallography, whereas others will rely on nuclear magnetic resonance spectroscopy to solve the structures. The challenge will be to take advantage of the diversity of approaches while avoiding redundancy, says John Norvell, director of the NIGMS project.

But some duplication of effort is desirable, as similar structures can arise from drastically different genetic sequences. It would also be helpful when groups are trying out new methods for discerning structures.

In addition, redundancy could come in the form of competition from similar projects. For example, Britain's Wellcome Trust is thinking about organizing an industrial structural genomics consortium (see *Nature* 406, 923; 2000).

But Stephen Burley, a Rockefeller University structural biologist, and one of the seven principal investigators selected by the NIGMS project, says that industrial researchers will be more interested in medically important proteins. The NIGMS project, in contrast, aims for breadth, not depth. ■

♦ <http://www.nigms.nih.gov/funding/psi.html>

US lab animals may win in lawsuit

Meredith Wadman, Washington

In a stinging defeat for the biomedical research lobby, the United States Department of Agriculture (USDA) is about to settle a lawsuit brought by animal-rights activists which seeks to bring some 23 million laboratory rats, mice and birds under the protection of the Animal Welfare Act.

If the suit is successful, 95% of laboratory animals will be subject to a new set of federal regulations. On Monday, the USDA was in the final stages of negotiating a settlement under which the agency would begin procedures that might lead to the extension of the law.

Advocates of biomedical research

criticized the move. "Settling this suit without taking into account the deep concerns of the research community is a serious mistake," said a statement from Jordan Cohen, the president of the Association of American Medical Colleges.

The lawsuit was filed last year by the Minnesota-based Alternatives Research and Development Foundation, which argued that the agriculture department had "arbitrarily and capriciously" excluded these animals from the 1966 law (see *Nature* 400, 197–198; 1999).

Researchers argue that other regulations already force them to take appropriate care of rats, mice and birds. ■

Age wins as Tokyo revamps retirement rules

David Cyranoski, Tokyo

The University of Tokyo is to extend the mandatory age of retirement from 60 to 65, despite criticism from some of its former professors. The increase will be phased in between 2001 and 2013, adding one year to the retirement age every three years.

The proposal has been hotly contested since it was put forward last December. Critics argue that it will enhance the power of established professors and block already congested career paths for young researchers. But the university president, Shigehiko Hasumi, says that the extension will help to move the university towards a "global standard" based on "diversity of nationality, sex and age".

Ken'ichi Arai, director of the university's Institute of Medical Science and a member of its senate, agrees. He says that flatly refusing to let those over 60 continue their work is age discrimination and a barrier to producing a university system "based purely on achievement".

Detractors allege, however, that the initiative is motivated not by lofty principles of equal opportunity, but by a desire to protect would-be retirees from declining post-

retirement employment prospects and changes in Japan's pension system.

In Japan, the pensionable age for public employees, including university professors, is also being raised from 60 to 65 — on the same 13-year schedule as at the university. In addition, the once-abundant opportunities for retired professors to augment a scanty pension with positions in private universities have been drying up.

Takashi Masuda, who retired from the university's faculty of sciences this March, believes that eliminating the need for retirees to seek employment elsewhere will make professors complacent. He has set up a webpage to air the opinions of faculty opposed to the change.

Extending the retirement age will make it more difficult for young researchers to get positions, Masuda says. This will be especially harmful in the sciences, he adds, where the most productive work is often done by scholars in their 30s or 40s.

Supporters and critics of the age change agree that its effectiveness will rest on new evaluation systems and other relevant policies that university departments must imple-

You can reach higher with that one—if you manage to get on...



ment as it comes into force.

Although optimists see the initiative as part of a larger move towards an achievement-orientated university system, others are less sanguine. According to Masuda, the change was passed by the university senate with no real debate, and little serious consideration will be given to improving the evaluation system, either. "The department heads won't want to make any hard decisions that they don't have to," he says.

♦ <http://www-masuda.cs.uec.ac.jp/~masuda/>

Frozen body offers chance to travel back in time

Quirin Schiermeier and Katrin Stehle

Six European research groups are eagerly awaiting new samples from the bones, teeth and intestine of Ötzi, the Stone Age man whose body was defrosted last week.

Ötzi's frozen body, over 5,000 years old, was found in 1991 in the Italian Alps close to the Austrian border. It was examined by local researchers for several years before being moved to a specially built museum of archaeology in Bolzano in Italy — under armed guard in case of attacks by Austrian nationalists (see *Nature* 391, 318; 1998).

Since then, tempers have cooled and a more coordinated approach has been established. A scientific committee set up in 1998 comprises researchers from Austria, Switzerland and Italy, and is chaired by Horst Seidler, co-director of the Institute of Anthropology at the University of Vienna.

Six projects have been selected for study. Scientists are particularly interested in analysing Ötzi's genetic make-up, if this proves technically possible. They hope to address medical, pathological and anthropological questions by examining bacterial DNA in his gut flora, to study genetic similarities with modern Europeans, and to learn about migration patterns and diseases in the late Neolithic period.

Previous examinations included isolation and analysis of mitochondrial DNA from tissue samples. But as only the surface of Ötzi's body was warmed up, data from many



The iceman cometh: Ötzi's fully defrosted body should yield high-quality samples for analysis.

samples were of low quality. Samples from the completely defrosted body should be more reliable and more readily comparable with other preserved human bodies.

In one of the projects, Franco Rollo, a professor of anthropology at the University of Camerino in Italy, plans to obtain bacterial genetic sequences from the iceman's intestine. Rollo hopes to determine Ötzi's eating habits and the distribution of pathogenic bacteria at the time he was alive.

Another research group, headed by Mark Thomas of the Centre for Genetic Anthropology at University College London, plans to investigate European migration patterns by comparing the iceman's Y-chromosome with genetic dispositions found among contemporary populations.

Another question concerns Ötzi's geographical origins. Wolfgang Müller, a researcher at the Institute for Isotope Geology at the Technical University of Switzerland in Zürich, hopes that the proportion of strontium and lead isotopes in Ötzi's tooth enamel will help provide the answer. If the quality of the enamel samples is high enough, Müller also wants to look into the oxygen isotopes they contain. This could help to determine the temperature conditions during Ötzi's early youth, he says.

JOSEF PERNTER/SÜDTIROLER ARCHÄOLOGIE-MUSEUM, BOZEN

Experts question precautionary approach

Colin Macilwain, Cambridge, Massachusetts

Eight months after the United Nations Biosafety Protocol became the first fully fledged international treaty to incorporate the 'precautionary principle', scientists and lawyers remain deeply divided on the principle's definition and merits.

Last week, at Harvard University's Center for International Development, experts gathered to discuss the impact of the protocol's precautionary approach on the growth of agricultural biotechnology. According to Calestous Juma of the Harvard centre, who convened the meeting, a lack of clarity is causing confusion in developing countries. "It is evident that there is no real agreement on what the precautionary principle means and how it should be applied," he says.

The principle states that potential environmental risks should be dealt with even in the absence of scientific certainty. It has long been advocated by environmentalists, who see it as a more effective way of managing hazards than traditional scientific risk assessments, which call for numbers and hard proof as prerequisites for action.

Some experts say that the principle is too vague to be of much value, either legally or scientifically. "It thrives because it is so ambiguous," says Gary Marchant, a law professor at Arizona State University. "What you end up with is arbitrariness." The sceptics say that the principle's impact on the spread of agricultural biotechnology around the world supports their view.

For Luiz Anthonio Barreto de Castro, head of genetic resources and biotechnology at EMBRAPA, the main agricultural research agency in Brazil, the application of the principle is of pressing concern. Transgenic crops have been banned in his country since 1998, he says, because of a judge's interpretation of a version of the principle included in the Rio Declaration, which emerged from the Earth Summit held in Brazil in 1992.

Yet in India, according to Aarti Gupta of Yale University, the principle has contributed to a well-balanced process in which extensive field trials are preceding the introduction of commercial transgenic crops. Indian regulators support the principle, she says, viewing sound science and a precautionary approach as synonymous.

John Mugabe of the African Centre for Technology Studies in Nairobi says that his continent has split into a hierarchy of approaches. Countries such as South Africa and Kenya are building regulatory regimes based on the principle. But poorer nations cannot do so, he says, and may instead accept transgenic crops simply on the basis that the United States considers them safe.

Critics of biotechnology claim that industrial lobbyists are trying to undermine the



Calestous Juma (left) argues that lack of clarity over the precautionary principle is causing confusion.

Biosafety Protocol. "The precautionary principle is now becoming enshrined in international law, and that is upsetting some powerful interests," says Philip Bereano of the University of Washington in Seattle. Some critics are even unhappy with Juma and Jeffrey Sachs, director of the Center for International

Development, for holding a seminar likely to highlight the principle's shortcomings.

Asked if he intended to praise the principle, or to bury it, Juma laughed and said: "We're a research institute. We're not interested in taking positions on the subject."

► <http://www.iisd.ca/sd/biotech/>

French lab seeks recipe for success

Declan Butler, Paris

Conventional culinary wisdom has it that a soufflé rises because of the dilation of air. In fact, it rises because water vapour is generated inside it — implying that the mould should be heated from the bottom, so that the rising vapour can push up through all the levels. A good crust helps too, as it prevents the vapour escaping.

Now Hervé This — the chemist who, with the late physicist Nicholas Kurti, explored soufflé thermodynamics a decade ago at the Clarendon Laboratory in Oxford — is to set up a laboratory of 'molecular gastronomy' at the Collège de France in Paris. The new facility will be in the Laboratory of Molecular Interaction run by Jean-Marie Lehn, the 1987 Nobel prizewinner in chemistry, who also has an interest in the field.

This's own long-standing interest has been in making cooking as much a science as an art. His PhD thesis was called 'Molecular and physical gastronomy', and the passion remained even after he left research in 1980 to join *Pour la Science*, the French edition of *Scientific American*, of which he later became editor. His return to research has been facilitated by a full-time position offered by INRA, the French national agricultural research organization.

The term 'molecular gastronomy' was coined in 1992 by This and Kurti. Now an international group of scientists who share an interest in the science of cookery plans to

establish it as a scientific field. The researchers already hold an annual international meeting on the subject in Erice, Sicily (see *Nature* 400, 17–18; 1999).

The group defines the field of molecular gastronomy as a distinct subdiscipline of food science. Its main goal is to test scientifically the cooking practices enshrined in the culinary wisdom that has resulted from centuries of empirical work by chefs.

There is no shortage of research questions. Hervé This and his colleagues intend to analyse the factors involved in the many kitchen activities leading to a successful dish, and invent new methods for preparing food. The results will be made available to chefs and to the public.

This himself has published several books, including the French best seller *Les Secrets de la Casserole*, and his frequent public lectures and television appearances have a large following.

He points out, for example, that microwave ovens often yield disappointing results because the temperatures obtained are too low to provoke the famous Maillard reaction between sugars and amino acids that generates many aromas and flavours.

But take some duck thighs, fry at a high temperature until brown, then take a syringe of Cointreau and inject it into the centre of the meat. Put it in a microwave oven for a few minutes and *voilà*: a perfect *canard à l'orange*, Maillard reaction included. ■

US Congress shows generosity towards physics projects

Washington The future of the National Ignition Facility, the troubled laser facility under construction at the Lawrence Livermore National Laboratory in California, looks more secure this week.

Congressional negotiators have ignored objections and approved \$199 million for it next year — just \$10 million less than the full amount that had been requested by President Bill Clinton's administration.

The negotiators also approved \$278 million for construction of the Spallation Neutron Source at the Oak Ridge National Laboratory in Tennessee — rejecting earlier threats to cut as much as \$150 million from the project. Congress was unexpectedly generous with funding for the Department of Energy, and other physics programmes received slightly more money than the administration had requested.

BSE committee delivers report to UK government

London Britain's BSE Inquiry committee handed over its report to the government on Monday. It documents the events that led

bovine spongiform encephalopathy to pass from cattle through the food chain to humans, becoming variant Creutzfeldt–Jakob disease (vCJD).

The Ministry of Agriculture, Fisheries and Food is expected to publish the report when parliament reconvenes on 23 October.

The inquiry, which began in December 1997 and has cost £27 million (US\$40 million), has taken evidence from 630 witnesses, including civil servants, agricultural and abattoir workers, scientists, physicians and family members of vCJD victims. The report arrives in 16 volumes, each focusing on a different aspect of the inquiry, with a short executive summary.

An inquiry spokeswoman said that it would provide “lessons for the future handling of potential public health issues, such as genetically modified foods”.

Kitt Peak telescope to reopen next year

San Diego A 1.3-metre telescope at Kitt Peak National Observatory near Tucson, Arizona, will be back in action next year after a five-year gap, thanks to a \$1 million refurbishment.

The telescope has been rebuilt with robotic controls that will allow its new operators — the Planetary Science Institute in Tucson,

Western Kentucky University and South Carolina State University — to make observations over the Internet.

There are also plans to save a second Kitt Peak telescope that was slated for closure in January by the National Optical Astronomy Observatory. The observatory says that a consortium will announce later this month that it will continue to operate the 0.9-metre telescope.

Buyer threatens to destroy stolen Enigma machine

London A man claiming to have innocently bought the Enigma coding machine stolen from the Bletchley Park Museum, Buckinghamshire, in April says he will destroy the device. The threat, made in a letter to the Bletchley Park Trust, follows the police's refusal to guarantee the man immunity from prosecution.

The trust had already received anonymous letters, seemingly from an intermediate, offering to return the machine if its buyer was reimbursed and not prosecuted (see *Nature* 407, 278; 2000). He is believed to have asked for around £25,000 (US\$37,000) for returning the missing encoder, a rare four-rotor version used by the Abwehr German military intelligence during the Second World War.

US company backs off basmati rice patent

New Delhi Dallas-based RiceTec, which caused a stir two years ago by obtaining a patent in the United States for basmati rice (see *Nature* 391, 728; 1998), has been forced to withdraw three of the 20 claims it made in the patent.

The change means that Indian basmati rice can now be sold in the United States without infringing the RiceTec patent, Indian officials say.

The withdrawal followed a suit filed by the Indian government to the US patent office in June. The company withdrew the claims in the light of 1,500 pages of evidence, provided by Indian scientists, that the grain characteristics purportedly developed by the company already exist in Indian basmati and are not unique.

Iran and US discuss science workshops

Washington The presidents of the US National Academy of Sciences, National Academy of Engineering and Institute of Medicine have met their Iranian counterparts in Tehran and promised to hold a series of further meetings.

Iranian and US scientists will discuss matters of mutual interest at six joint workshops, to be held over the next two years. Although

topics have not been determined, they are likely to cover marine ecology in the Caspian Sea and the Persian Gulf, urban pollution in Tehran, scientific ethics, and the 'brain drain' of scientists and engineers from Iran, according to a statement from the US academies.

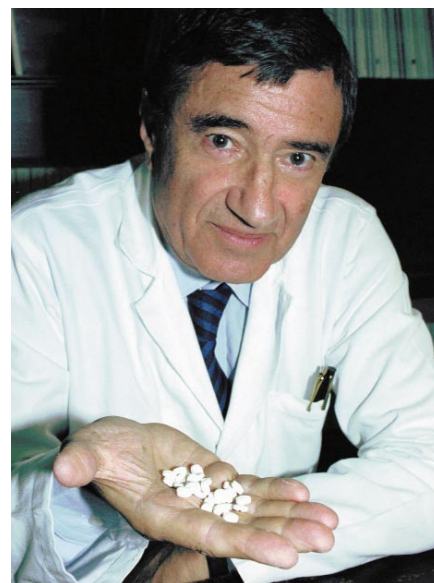
The US scientific leaders also toured six universities and other research facilities in Tehran and Isfahan. Their visit followed one to Washington by the leaders of the Academy of Sciences and the Academy of Medical Sciences of the Islamic Republic of Iran, in September 1999.

Abortion pill approved for use in the US

Washington The US Food and Drug Administration (FDA) has finally given the abortion pill mifepristone, better known as RU-486, approval for use in the United States.

The clearance comes 20 years after the compound was discovered in France and 17 years after the FDA first issued a testing permit for the drug. The decision is a major setback to the powerful anti-abortion lobby in the United States.

Marketed as Mifeprex, the pill will be available to doctors in about a month's time. It can be prescribed to women who are less than seven weeks pregnant, although it must



Emile-Etienne Beaulieu, discoverer of RU-486: it will be available in the US in about a month.

be taken in combination with the synthetic prostaglandin misoprostol.

The Population Council, a non-profit organization, holds the US patent for the drug, which will be marketed by Danco Laboratories of New York. The FDA and the Population Council have kept the name of the pill's US manufacturer private.

The sandman's secrets

Neuroscientists at last have a molecular handle on how the brain controls our daily cycle of sleep and wakefulness. This is opening a rich avenue of study that could lead to new therapies for sleep disorders, says Marina Chicurel.

Let sleeping dogs lie, warns the old proverb. But Emmanuel Mignot has helped to open up an exciting area of molecular neuroscience by ignoring this advice. In 1989, Mignot and his colleagues at Stanford University School of Medicine in California began to study a group of Doberman pinschers afflicted by narcolepsy — a disorder marked by uncontrollable urges to sleep and sudden bouts of muscle weakness. They hoped that the dogs would provide a suitable animal model in which to uncover the causes of this distressing human disease.

More than a decade later, that hope is being realized. But, more significantly, Mignot's project has also converged with several other lines of research to suggest that a pair of neuropeptides called the hypocretins/orexins are central players in controlling the body's cycle of sleep and wakefulness. As such, the peptides should help fill an important knowledge gap.

Biologists have made great strides in understanding the molecular components of the internal clocks that control our daily, or circadian, biological rhythms. But they have struggled to reveal how these clocks influence our sleep patterns. "The link between circadian rhythms and sleep is one of the great unsolved problems," says Allan Hobson, a leading sleep researcher at Harvard Medical School in Boston. "These neuropeptides are the first specific candidates."

With millions of people suffering from sleep disorders, the pharmaceuticals industry is interested in the hypocretins/orexins — better sleeping pills, drugs to combat drowsiness and treatments for conditions

such as narcolepsy could emerge from research on the neuropeptides.

The peptides should also shed light on other elements of the body's 'housekeeping' processes — such as the control of appetite and blood pressure. "It's an amazing culmination of a series of leads that have come together over a relatively short period of time," says Thomas Kilduff, a neuroscientist at SRI International, a non-profit research corporation in Menlo Park, California.

Clocking on

One of the earliest leads emerged from Gregor Sutcliffe's lab at the Scripps Research Institute in La Jolla, California. In 1994, he started looking for molecules that regulate basic body functions such as eating, sleeping and reproduction. Neuroscientists had long known that a region in the brain called the hypothalamus controlled most of these processes by producing specific peptides to transmit signals within the brain and to other organs. But the number of known neuropeptides was small compared with the range of activities governed by the hypothalamus. "We argued there must be a lot of hardware left to discover," says Sutcliffe.

So Sutcliffe and his colleagues searched for genes that were expressed only in the hypothalamus. Using messenger RNA (mRNA) extracted from rat hypothalamus tissue, the researchers generated a library of single-stranded complementary DNA (cDNA). To separate out genes expressed in the brain in general from those specific to the hypothalamus, the workers mixed all the fragments with mRNA from two other brain regions. Many of the cDNAs paired up with these mRNAs, forming double-stranded molecules. But 38 of the fragments could find no partner and remained single-stranded.

One in particular, dubbed 'clone 35', attracted the researchers' attention because its mRNA was found exclusively within a cluster of cells in the lateral hypothalamus¹. Because such cell clusters tend to regulate specific functions, clone 35 seemed to fit the bill for the type of regulatory hardware the researchers were looking for. "As soon as we saw that clone," recalls Kilduff, a member of the research team, "we knew we had struck gold." Sutcliffe and his colleagues predicted

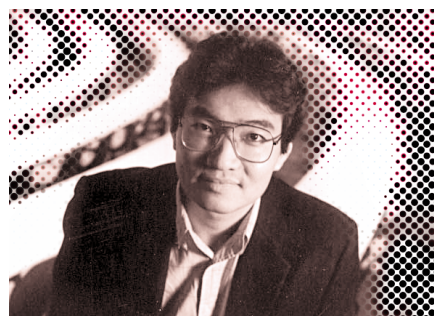


Dog tired: studies of narcoleptic Doberman pinschers have linked the control of sleep to two peptides that are thought to regulate appetite.

the structure of two peptides encoded by clone 35 and named them hypocretins, because they were produced in the hypothalamus and resembled the incretin neuropeptides, which help regulate gut function.

Meanwhile, at the University of Texas Southwestern Medical Center at Dallas, Masashi Yanagisawa and his colleagues were rummaging through a mass of 'orphan' receptor molecules. Found on the surface of cells, these molecules transmit a signal to the cell's interior when they bind to neuropeptides or other transmitter molecules. Genomics studies had identified over 100 human DNA sequences as coding for surface receptors coupled to G proteins, which are important in cell signal transduction. But most of these receptors had no known partner. Yanagisawa reasoned that if he could find some of the transmitters that bound to them, he would gain insights into brain function — with possible medical applications. "Sixty per cent of clinically used drugs target one or more of the G protein-coupled receptors," he notes.

Sifting through brain extracts, Yanagisawa's team discovered two peptides, derived from a single precursor protein, that stimulated two closely related orphan receptors². Again, the peptides were produced in the



Yanagisawa: infrared snooping on knockout mice offered clues to the causes of narcolepsy.



lateral hypothalamus. And like Sutcliffe, Yanagisawa realized that this might mean that they were involved in regulating basic body functions. Because previous work had linked the lateral hypothalamus with food intake, he suspected feeding behaviour in particular. And when his team injected the peptides into the brains of rats, the animals did indeed begin to eat more. So Yanagisawa named the two peptides orexins A and B after the Greek word *orexis*, meaning appetite.

Sutcliffe and Yanagisawa only realized the overlap between their work after they had published their papers, just six weeks apart. Sutcliffe's hypocretin mRNA, it turned out, encoded the two orexins. But it took another coincidence — this time involving a convergence between Yanagisawa's work and Mignot's narcoleptic dog project — to make the connection with sleep.

Unlikely bedfellows

Mignot had been struggling to identify genes associated with narcolepsy for the best part of a decade. But in 1997, the pace picked up after his team and other collaborators built a library containing large DNA fragments from the genome of a dog that was heterozygous for a narcolepsy gene — carrying one copy of the normal gene and one of the mutated, disease-causing version. In less than two years, the team narrowed the search for the gene to a genomic region containing two

candidates. One was not expressed in the brain, but the other encoded one of the two orphan receptors studied by Yanagisawa, the hypocretin/orexin 'type 2' receptor.

"It was a good candidate," says Mignot. "But at the same time we weren't so sure." After years of painstaking effort, and one previous false alarm³, Mignot was reluctant to focus all of his efforts on a single receptor. Also, he could not quite believe that the gene for which he had long been searching encoded a receptor for neuropeptides that a former labmate had just helped discover — Kilduff had worked with Mignot at Stanford before teaming up with Sutcliffe. "It was such a coincidence, it was impossible," says Mignot, who now reckons that he could have ended his quest six months earlier, had he ignored his disbelief and followed the lead.

By the time Mignot had dispelled his doubts, Yanagisawa's group had created 'knockout' mice in which the genes encoding the two neuropeptides were disabled. Convinced that the peptides primarily regulated appetite, Yanagisawa and his colleagues probed their mice for feeding abnormalities, taking the unusual step of using an infrared camera to monitor the animals' nocturnal eating habits. This night-time snooping revealed that the mice would suddenly collapse, often while walking or grooming. Suspecting that the animals were suffering from seizures, the researchers used electrodes to monitor the rodents' brainwaves. They found that, far from experiencing seizures, the mice were falling asleep. "These mice fulfilled all the objective criteria of human narcolepsy," says Yanagisawa. In August last year, only two weeks after Mignot published his findings on narcoleptic dogs⁴, Yanagisawa reported his knockout-mouse study⁵.

Since then, the link between the neuropeptides and human narcolepsy has strengthened. First, Mignot and his col-

leagues measured the levels of one of the neuropeptides, hypocretin 1/orexin-A, in the cerebrospinal fluid of nine narcoleptic patients. Seven had levels below the detection limit⁶. And just last month, Mignot's group and a team led by Jerome Siegel at the University of California at Los Angeles published separate papers revealing that post-mortem brains of human narcoleptics contain almost no hypocretins/orexins^{7,8}. But unlike Mignot's dogs, human narcoleptics only very rarely seem to have mutated hypocretin/orexin genes⁷. Instead, it seems that they have lost most, if not all, of the cells that produce the neuropeptides. Siegel found scar tissue in the area normally occupied by the hypocretin/orexin-producing cells⁸, consistent with the idea that narcolepsy might be an autoimmune disorder^{9,10}.

Dreams of drugs

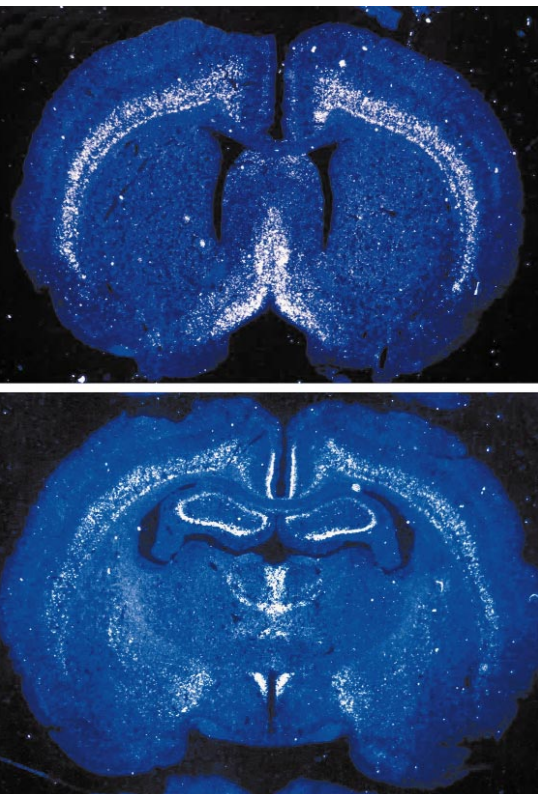
Neuroanatomical studies have reinforced the idea that the neuropeptides help control the sleep-wake cycle¹¹. "All the important structures in sleep regulation receive a hypocretin/orexin input," says Christelle Peyron, who recently moved from Stanford to the Geneva University Hospitals, and who first described the distribution of the hypocretins/orexins in the brain. This has led researchers to speculate that research on the peptides could result in new drugs to treat a variety of sleep disorders. Drugs that mimic the neuropeptides, for instance, might do for narcoleptics what insulin does for diabetics. And if the basic role of the hypocretins/orexins is to maintain wakefulness or act as a 'gatekeeper' of sleep, they might also combat normal sleepiness — such as that experienced by night-shift workers. Conversely, molecules that block the neuropeptides' effects might be used to develop improved sleeping pills.

But many hurdles lie ahead. "The pharmacology of the peptides is in its infancy," says Neil Upton of SmithKline Beecham in Harlow, Essex, which holds commercial rights to aspects of Yanagisawa's work. Treating narcolepsy using hypocretin/orexin mimics is the most immediate prospect. But Upton and his colleagues have already generated a small molecule that crosses the blood-brain barrier and binds to the type 1 hypocretin/orexin receptor, preventing it from being stimulated by the neuropeptides. Hypocretins/orexins increase arousal when injected into the brains of rats¹², but SmithKline Beecham's molecule seems to reverse this effect. The team has yet to assess whether the molecule will promote sleep in normal rats that have not received hypocretin/orexin injections, and need to design more stable molecules that will persist in the brain for several hours.

Another problem is that tinkering with the hypocretin/orexin system could do more than just influence sleep. SmithKline Beecham's molecule reduces appetite — an effect for



Kilduff: sleep research has struck gold.



In control? Sections of rat brain showing cells bearing the hypocretin/orexin receptors (white).

which it was, in fact, originally designed. Unpublished results obtained by Takeshi Sakurai of the University of Tsukuba in Japan further complicate the picture. He has developed a mouse model of narcolepsy in which, like human narcoleptics, the animals lack hypocretin/orexin-producing cells. These mice gain weight, despite eating 20% less than normal mice. Sakurai speculates that a lack of the peptides reduces metabolic rate.

Positive signals

Other researchers have reported that the neuropeptides affect functions such as cardiovascular activity^{13,14}, and hormone secretion^{15,16}. Teasing apart these multiple functions will probably require neuroscientists to elucidate the distinct roles of the two hypocretin/orexin receptors. Another complication is that the peptides not only transmit signals directly to other neurons, but also act as 'neuromodulators', influencing the release of other neurotransmitters¹⁵.

Several groups are trying to dissect the neuropeptides' effects on different brain regions. Working with rats, for example, Luis de Lecea at Scripps has injected the peptides into the locus coeruleus. This area of the brain receives the lion's share of the connections from hypocretin/orexin-producing cells, and is implicated in promoting alertness and maintaining muscle tone. He observed an increase in wakefulness¹⁷. In similar, unpublished, work, Siegel noted an increase in muscle tone.

The neuropeptides offer an unprecedented opportunity to study how the brain divides our lives between sleep and wakefulness.

But the locus coeruleus is not the sole, or even necessarily the primary, mediator of the peptides' effects on sleep and wakefulness. For a start, it sports only type 1 hypocretin/orexin receptors, not type 2. In addition, experiments that flood brain areas with excess hypocretins/orexins suffer from the limitation that the added molecules can diffuse away from the target area and affect neighbouring regions. Yanagisawa, who has developed two lines of knockout mice, each lacking one of the two hypocretin/orexin receptors, is beginning to probe the receptors' distinct functions. So far, his unpublished results indicate that, although the type 1 receptor contributes to sleep regulation, the type 2 receptor plays a more central role.

Rousing research

Trying to integrate the recent hypocretin/orexin findings with decades of sleep research, some neuroscientists are proposing models to help guide future experiments. Kilduff and Peyron, for example, have sketched out a model¹⁸ in which hypocretin/orexin cells promote wakefulness by stimulating several brain regions, including the locus coeruleus. They suggest that the hypocretin/orexin cells are active both during waking and REM (rapid eye-movement) sleep — based on electrophysiological studies showing that some cells in the lateral hypothalamus display this pattern of activity. But during non-REM sleep, which normally comprises the transition between waking and REM sleep, a region of the brain called the preoptic area inhibits the hypocretin/orexin cells.

As an organism enters REM sleep, Kilduff and Peyron suggest, this inhibition dampens and the hypocretin/orexin cells fire up again. Although the researchers propose that hypocretin/orexin cells are active both during waking and REM sleep, they believe that the neuropeptides have different effects during these two phases because of the modulatory activity of other brain regions. For example, cells in the locus coeruleus would be active during waking because of their stimulation by the hypocretin/orexin cells, but would quieten down during REM sleep because of inhibitory inputs from other brain structures.

Not everyone agrees with this model. "If you follow that model through, the lack of hypocretins/orexins should cause continuous sleepiness," says Clifford Saper at the Beth Israel Deaconess Medical Center and Harvard Medical School in Boston. "This is not really the case." Saper thinks that Kilduff

and Peyron's model is missing a negative feedback loop that would ensure organisms go quickly from sleeping to waking, and vice versa. "You can visualize this as a see-saw, with either one side up, or the other side up, but spending very little time in transition," says Saper. "The hypocretin/orexin input acts like a finger pushing down on the side that favours wakefulness. Without it, the system becomes unstable, with people or animals flipping out into sleep states abruptly, which is the hallmark of narcolepsy."

Testing any model will involve monitoring the activity of hypocretin/orexin-producing cells during sleep and wakefulness. Thomas Scammell, also at the Beth Israel Deaconess Medical Center and Harvard, has taken a first step by monitoring the expression of *Fos*, a gene used as a marker of neuronal activity. His team's unpublished findings suggest that hypocretin/orexin cells seem to be most active during wakefulness, but are quiet during REM sleep, suggesting another revision to Kilduff's model. But ultimately, electrophysiological recordings will be needed to characterize the hypocretin/orexin cells' behaviour more precisely.

Although the precise mechanics of hypocretin/orexin function remain to be elucidated, neuroscientists agree that the neuropeptides offer an unprecedented opportunity to study how the brain divides our lives between sleep and wakefulness. It is a far cry from just three years ago, when Kilduff got a less than enthusiastic reaction when he talked at meetings about a neuropeptide system of then unknown function. Today, there is little chance of anyone in the audience dozing off while Kilduff or any of the other major players in hypocretin/orexin research gives a presentation. "On the other hand," jokes Kilduff, "as long as I see people doing that, I know I'll have a job."

Marina Chircurel is a writer in Santa Cruz, California.

1. de Lecea, L. *et al. Proc. Natl Acad. Sci. USA* **95**, 322–327 (1998).
2. Sakurai, T. *et al. Cell* **92**, 573–585 (1998).
3. Mignot, E. *et al. Proc. Natl Acad. Sci. USA* **88**, 3475–3478 (1991).
4. Lin, L. *et al. Cell* **98**, 365–376 (1999).
5. Chemelli, R. M. *et al. Cell* **98**, 437–451 (1999).
6. Nishino, S. *et al. Lancet* **355**, 39–40 (2000).
7. Peyron, C. *et al. Nature Med.* **6**, 991–997 (2000).
8. Thannikail, T. C. *et al. Neuron* **27**, 469–474 (2000).
9. Juji, T. *et al. Tissue Antigens* **24**, 316–319 (1984).
10. Mignot, E. *Adv. Neuroimmunol.* **5**, 23–37 (1995).
11. Peyron, C. *et al. J. Neurosci.* **18**, 9996–10015 (1998).
12. Piper, D. C., Upton, N., Smith, M. I. & Hunter, A. J. *Eur. J. Neurosci.* **12**, 726–730 (2000).
13. Samson, W. K. *et al. Brain Res.* **831**, 248–253 (1999).
14. Shirasaka, T. *et al. Am. J. Physiol.* **277**, R1780–R1785 (1999).
15. van den Pol, A. N., Gao, X. B., Obrietan, K., Kilduff, T. S. & Belousov, A. B. *J. Neurosci.* **18**, 7962–7971 (1998).
16. Pu, S. *et al. Regul. Pept.* **78**, 133–136 (1998).
17. Bourgin, T. *et al. J. Neurosci.* (in the press).
18. Kilduff, T. S. & Peyron, C. *Trends Neurosci.* **23**, 359–365 (2000).

A hundred million points of light

Covering a quarter of the sky, the Sloan Digital Sky Survey has a bright future as a key astronomical resource, says Govert Schilling.

Every day, a sliver of the Universe travels from the Apache Point Observatory in New Mexico to the Fermi National Accelerator Laboratory near Chicago. By express courier, a computer tape bearing one night's observations made by the Sloan Digital Sky Survey's 2.5-metre telescope is ferried to Fermilab's Feynman Computing Center. In five years time, these observations will cover a quarter of the sky and will have been collected into one of the largest scientific databases ever produced — totalling some 15 terabytes (15 million megabytes) of data.

Already, astronomers are aware of the survey's importance for studying faint objects such as distant quasars, and for examining the large-scale structure of the Universe. But the project's leaders believe that the field has yet to grasp the survey's wider significance. According to team member Bruce Margon of the University of Washington in Seattle, the Sloan survey will mark a shift towards a new way of doing astronomy. For many projects, he claims, it will no longer be necessary to book time on leading telescopes. "You will never leave your desk, but just tap into the Sloan archive," he says.

The Sloan Digital Sky Survey began gathering data in 1998, but it will be officially dedicated in a ceremony at Apache Point on 6 October. The survey is valuable because it will record data on the brightness of more than 100 million celestial objects at five different wavelengths, or colours. This should give researchers in just about every subfield of astronomy a wealth of data to work with. "Sloan includes every type of observation an astronomer might want to do," argues Margon. "This is the first time we have had a digital archive with such capability."

The project will also use spectroscopy to record the redshifts — and so work out the distances from the Earth — of one million galaxies, providing a three-dimensional map of our cosmic neighbourhood. "They have 20,000-



Star tracking: after dark, Apache Point's 2.5-metre telescope emerges from its housing (inset) to scan the sky.

plus in the can already," says John Huchra of the Harvard-Smithsonian Center for Astrophysics in Cambridge, Massachusetts.

This spectroscopic survey was the original motivation for the project. It will let astronomers study the Universe's large-scale structure by mapping the distribution of the galaxies. But the survey's photometric data might also play a role. For distant galaxies, instead of using time-consuming spectroscopy to measure redshifts, astronomers will be able to generate 'photometric redshifts' from Sloan's five-colour brightness data. Although this could result in galaxies with unusual colours being assigned incorrect redshifts, it offers a rapid way of getting redshifts for large numbers of galaxies.

So far, many of the Sloan survey's highest-profile discoveries have come from expected quarters, such as the study of distant quasars — very remote galaxies with bright cores, probably powered by massive black holes. Michael Turner of Fermilab, the scientific spokesman for the survey, boasts that eight of the ten most distant quasars known — including the current record-holder¹ — were bagged by Sloan. "We've found 2,000 quasars so far," he adds.

The survey is also yielding a rich haul of solitary brown dwarfs, faint objects that are too massive to be classed as planets but not sufficiently large to burn as stars^{2,3}.

But Bob Hanisch of the Space Telescope Science Institute in Baltimore predicts that the most important scientific results will come from unexpected angles, as astronomers start playing around with the Sloan survey data. "Large-scale statistical studies will point out new classes of objects that can be studied in more detail with large telescopes," says Hanisch. "We're on the threshold of a new revolution that will have a very broad impact."

Piero Benvenuti of the European Southern Observatory in Garching, Germany, agrees. The Sloan is the most ambitious of several survey and data-archiving projects, operating at a variety of wavelengths, he says. If these can be merged and extended, astronomers could mine 'old' data for most of their work, and state-of-the-art telescopes could be devoted exclusively to projects that need their unique capabilities.

US astronomers now want to build a National Virtual Observatory incorporating the Sloan and other surveys. And in the long-term, this might become a truly global initiative. The reward, predicts Benvenuti, will be "a giant leap in efficiency".

Govert Schilling is an astronomy writer in Utrecht.

1. Fan, X. *et al.* *Astron. J.* **120**, 1167–1174 (2000).
2. Strauss, M. A. *et al.* *Astrophys. J. Lett.* **522**, L61–L64 (1999).
3. Tsvetanov, Z. I. *et al.* *Astrophys. J. Lett.* **531**, L61–L65 (2000).

http://www.sdss.org

Legal confusion over 'cloning' risks throwing baby out with bathwater

Sir—The UK Chief Medical Officer has supported the recommendation¹ that the 1990 Human Fertilisation and Embryology (HFE) Act should be amended to allow therapeutic cloning of humans, for research purposes only². He also recommends: "The transfer of an embryo created by cell nuclear replacement into the uterus of a woman (so-called 'reproductive cloning') should remain a criminal offence" (recommendation 7 in ref. 3).

This recommendation implies that cell nuclear replacement and cloning are one and the same. They are not. If the recommendation is accepted as it stands, it risks perpetuating a confusion in UK law⁴—already found in the HFE Act: section 3 (3) (d)—and possibly thereby influencing laws elsewhere. If so, this might reduce the future treatment possibilities for some infertile couples.

Nuclear transfer constitutes reproductive cloning only when the individual so created is genetically identical to the nuclear donor. However, it is theoretically possible to use the nuclear transfer technique to generate a genetically biparental child. For couples in which one or both completely lack any germ cells or the means of producing them, the only current treatment option is gamete or embryo donation. However, one or both parents then lack a genetic contribution to their offspring. In the future, such couples may be able to overcome this problem by use of a nuclear transfer approach in which haploid nuclei are generated from parental diploid somatic cells and not via gametes.

Two types of approach might be used to achieve this outcome. In the first approach, a somatic cell from one parent or from each parent would be induced *in vitro* to undergo meiotic reduction to produce two haploid cells, nuclei from which would then be used as donors for transfer to a recipient oocyte. A second approach would be to generate a tetraploid egg by transfer of one diploid somatic nucleus from each parent to an enucleated pre-anaphase I oocyte capable of undergoing a reduction division and then subsequent activation, so as to establish diploidy.

Although several biological and technical problems must be solved before these approaches are available and safe, such therapy does not seem beyond the realm of eventual possibility. However, both these approaches use nuclear transfer and so would be prohibited under UK law.

We draw attention to these possibilities

now, both to encourage discussion of their social desirability and ethical status, and to stimulate a more accurate and informed use of language in science, politics and law.

If legislation is being proposed, it must say what it means. If the legal intention is to prohibit production of an individual who is genetically identical or almost identical to another individual, then the law should use words that attain this objective, and this only. If it does not do so, potentially beneficial applications of nuclear transfer technology for human reproductive purposes may in future be prohibited accidentally.

Martin H. Johnson*, **Jacek Z. Kubiak†**

*Department of Anatomy, Downing Street, Cambridge CB2 3DY, UK

†Department of Biology and Genetics of Development, UPR 41 CNRS, University of Rennes 1, Faculty of Medicine, 2 avenue du Professeur Léon Bernard, CS 34317, 35043 Rennes cedex, France

1. HGAC/HFEA *Cloning Issues in Reproduction, Science and Medicine* (The Human Genetics Advisory Commission and The Human Fertilisation and Embryology Authority, London, 1998). <http://www.dti.gov.uk/hgac/>
2. Dickson, D. & Smaglik, P. *Nature* **406**, 815 (2000).
3. Donaldson, L. *Stem Cell Research: Medical Progress with Responsibility* (Department of Health, London, 2000). <http://www.doh.gov.uk/cegc/stemcellreport.htm>
4. Human Fertilisation and Embryology Act (HMSO, London, 1990).

'Benign neglect' of inner city led to TB epidemic

Sir—Thomas Dormandy's review and Richard Coker's book *From Chaos to Coercion: Detention and the Control of Tuberculosis* (*Nature* **405**, 995–996; 2000), about the late 1980s epidemic in New York City, both mention the socioeconomic context of tuberculosis (TB): "poverty, overcrowding, ignorance, homelessness, drug abuse and social alienation". But neither discusses the history of this particular epidemic.

The federal and local government response to inner-city problems in the 1970s was to try to clear out these communities with policies of 'benign neglect' and 'planned shrinkage'. This meant reducing essential services in areas of need. More than 10% of the fire-fighting units in poor minority areas of New York City were cut. This led to an epidemic of fires, destroying between 200,000 and 300,000 poor and working-class homes.

An estimated 600,000 poor people were displaced by the housing loss: the 1980 census showed that 1.3 million white people left New York between 1970 and 1980. This flight was their response to the deteriorating physical and social conditions caused by the burn-out. Among those who remained, the incidence of TB increased dramatically. TB is linked

with overcrowded conditions, social marginalization, substance abuse and poor nutrition—all factors influenced by this forced migration. The increase was sharpest in communities with large populations of the precariously housed (people moving in to share rented accommodation with the existing tenants).

Inner-city neglect and loss of fire-fighting units did not happen by chance. If they had not been carried out as public policy, there would have been no New York TB epidemic and no need for coercion and detention of patients. The lack of regard for the populations targeted by 'benign neglect' and 'planned shrinkage' is clear from the length of time the epidemic was allowed to rage through these communities. Only when it spilt over into middle-class neighbourhoods and into the suburbs did a serious control effort begin.

Deborah Wallace

Department of Sociomedical Sciences, Joseph L. Mailman School of Public Health, Columbia University. Home address: 549 W. 123 Street, #16F, New York, New York 10027, USA

Loss of taxonomists is a threat to pest control

Sir—The cover of the issue reporting the landmark genome sequencing of the bacterium *Xylella fastidiosa* (*Nature* **406**, 151–157; 2000) shows a vector, but we are not told what sort of insect it is. Michael Bevan's accompanying News and Views article (*Nature* **406**, 140; 2000) reveals that the vectors are xylem-feeding leafhoppers, but no names are mentioned.

The culprit is, in fact, a species of *Acrogonia*, a member of the tribe Proconiini of the subfamily Cicadellinae. All members of this leafhopper subfamily are xylem feeders, and some species in several genera are known or potential vectors of both the citrus disease and Pierce's disease of vines in the southern United States. It is not possible to identify the *Acrogonia* species from the cover photograph: colleagues in Brazil confirm that it could be an undescribed species. There is much taxonomic work to be completed.

Insect taxonomists are a threatened species worldwide. Have we already reached the situation where more scientists can sequence a genome than can identify the potential vectors? Control measures against *Xylella* and other similar organisms will depend not only on such advances but also on being able to elucidate the biology of potential vectors once they have been identified by taxonomists.

M. R. Wilson

Department of Biodiversity & Systematic Biology, National Museum of Wales, Cardiff CF10 3NP, UK

Research, innovation and politics

What future for science and technology after this autumn's US elections?

**David M. Hart and
Lewis M. Branscomb**

The 7 November election in the United States promises to be one of the most closely fought in decades. The presidency is the biggest prize, but it is also possible that one or both houses of Congress will change hands. Moreover, with two or more of the nine Supreme Court justices likely to step down during the next four years, a change in the country's third branch of government could follow.

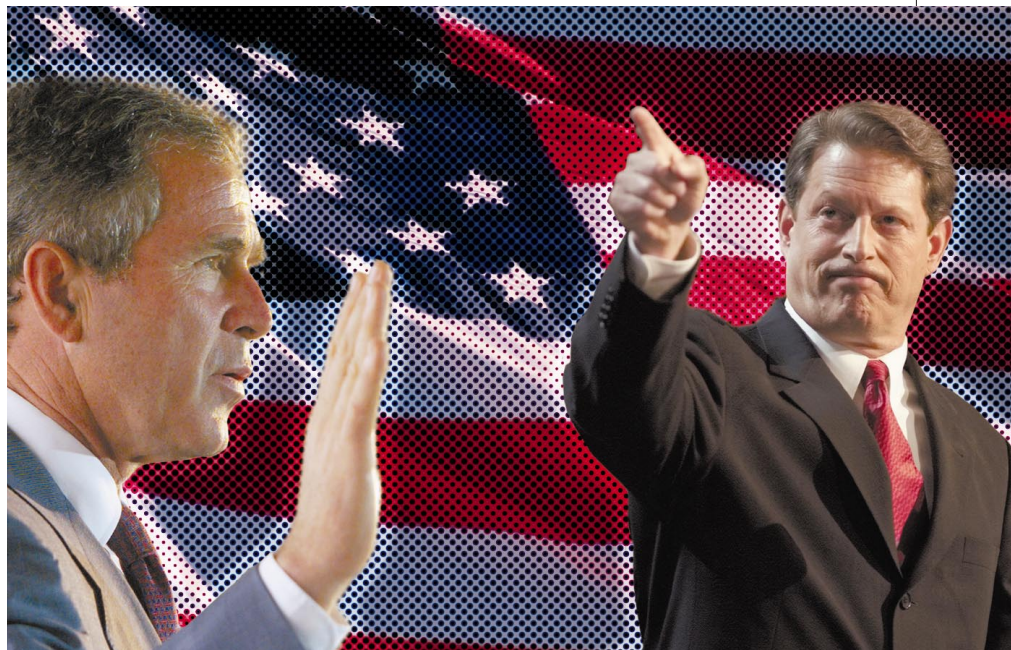
Although the two main political parties share a vision of an economy and society built on science and technology that is open to the rest of the world, they have fundamental differences about the government's role in research and innovation. Scientists and those with an interest in science everywhere have much at stake in the election's outcome.

Parties' choice

Judged by the level of R&D spending from all sources, the state of US research and innovation is extremely healthy. According to the National Science Foundation, total spending on R&D in the United States started accelerating in 1995 and has grown at an annual rate of 5% or more in real terms ever since — including an estimated 7% in 1999. Corporate R&D spending has been particularly strong; the industrial sector in 1999 accounted for \$169 billion, nearly 70% of the national total of \$247 billion (see Fig. 1, next page).

Both parties applaud the growth of corporate research, but they diverge over the relationship between the public and private sectors. Vice-president Al Gore, the Democratic presidential candidate, has been an avid advocate of public-private partnerships in research and innovation, which he helped to make a centrepiece of President Bill Clinton's policy. Republicans in Congress have sought a more rigid separation of the public and private spheres, attributing the credit for the present prosperity to entrepreneurship rather than partnership, a theme echoed by Texas governor George W. Bush, the Republican presidential standard-bearer.

Although federal appropriations comprise a shrinking minority of the nation's R&D (some \$80 billion out of the \$247 billion total in 1999), the government remains the dominant source of academic research support (\$16.1 billion out of \$28.3 billion) and is a critical agenda-setter in many applied fields. From fiscal year 1996, when the federal R&D budget was about \$75 billion (in inflation-adjusted 2000 dollars, according to the American Association for



the Advancement of Science), to fiscal year 2000, when it reached \$83 billion, the federal R&D budget grew by about 10% in real terms, reversing a decline of about the same magnitude in the previous four years. This change has everything to do with the turnaround from deficit to surplus in the larger budget picture — R&D comprises a sufficiently substantial share of the aggregate federal budget (about 15% of discretionary spending) that it rises and falls with the aggregate.

National economic growth has been extremely strong in recent years. One's view of the two parties over the scale of federal R&D spending, then, ought to turn on one's assessment of their competing fiscal plans: the Republicans stress tax cuts and greater reliance on markets; the Democrats focus on debt retirement and greater reliance on public appropriations.

Health and defence

The composition of federal R&D spending is a different matter. About three-quarters of the decline in total federal R&D spending between fiscal years 1992 and 1996 was accounted for by cuts in the Department of Defense (DoD) — from \$43.7 billion inflation-adjusted 2000 dollars in fiscal year 1992 to \$37.9 billion in 1996. By contrast, 63% of the increase between fiscal years 1996 and 2000 was for the National Institutes of Health (NIH), from \$12.1 billion inflation-adjusted 2000 dollars in fiscal year 1996 to \$17.1 billion in 2000.

Candidates are facing science and technology issues that do not divide along traditional party lines

Both these trends had bipartisan support, but the two parties appear to differ when looking to the future. In recent years, Congressional Republicans have pushed to increase the DoD budget above the administration's request, and pressed for early deployment of a national missile defence system. Bush has taken up both these themes, whereas neither Clinton nor Gore has made expansion of defence R&D a high priority.

If the rhetoric of the two campaigns is to be believed, there should be a continued healthy growth in the NIH budget in the next four years. Gore has promised to double federal biomedical research spending, and Bush has matched that. It is difficult to evaluate how realistic these proposals might be, as both extend over the better part of the next decade. The two camps have squabbled over each other's figures, but the important differences lie in the specifics.

Perhaps the most important fight has to do with stem-cell research. The Clinton administration has enraged some Congressional Republicans by promulgating regula-

CORBIS

tions that would permit federally funded researchers to use human stem cells derived from embryos under certain conditions (see *Nature* 406, 815; 2000). Bush's supporters in the National Right to Life Committee say he would overturn the new policy if elected.

Other priorities

The rest of the federal R&D budget has hardly been mentioned in the current campaign, but the recent past might provide some clues about what may happen after the election. The National Science Foundation (NSF), for instance, was budgeted for an increase of nearly 20% in fiscal year 2001. But this thrust towards a civilian R&D portfolio more balanced between medical and non-medical fields has not fared well in the Republican-dominated House of Representatives, which agreed to just a 4% rise.

The Clinton administration's proposed budget for NASA, too, has been trimmed by about 3% by the House, although both parties have criticized the agency in the wake of its recent disappointments. The civilian R&D elements of the Department of Energy (DoE) reveal the same pattern even more sharply, with some Congressional Republicans calling for this \$7 billion R&D agency to be disbanded altogether. (Much of the R&D spending in the unlikely eventuality of DoE disbandment would almost certainly be re-allocated to other government departments.) The House voted a 5% cut in the administration's request for DoE R&D, which the Senate has now restored.

The budget endgame for all these agencies, in which the Senate will take its action on the rest of the appropriations bills and the two houses will confer with one another and with the White House, is in progress as this issue goes to press. The final figures for the upcoming fiscal year, which begins on 1 October, are likely to be closer to the administration's proposed budget than the House of Representatives' initial appropriations.

One source of Republican dissatisfaction with the DoE is the agency's leading role in US policy on global climate change, with which Gore is closely associated. Gore spoke of this concern in his convention address, and is committed at least to the Kyoto Protocol on climate change and perhaps to even stronger measures to limit the growth of greenhouse-gas emissions. Scepticism among Republicans has discouraged the Clinton administration from submitting the Kyoto Protocol to the Senate for ratification. Bush, a former oil-industry executive, opposes the treaty, and his running mate, former Defense Secretary Dick Cheney, spoke out against it when chief executive officer at oil services giant Halliburton.

The parties' differences on global climate-change policy reflect a general tendency of the Republicans to place a higher burden of proof than the Democrats on scientific

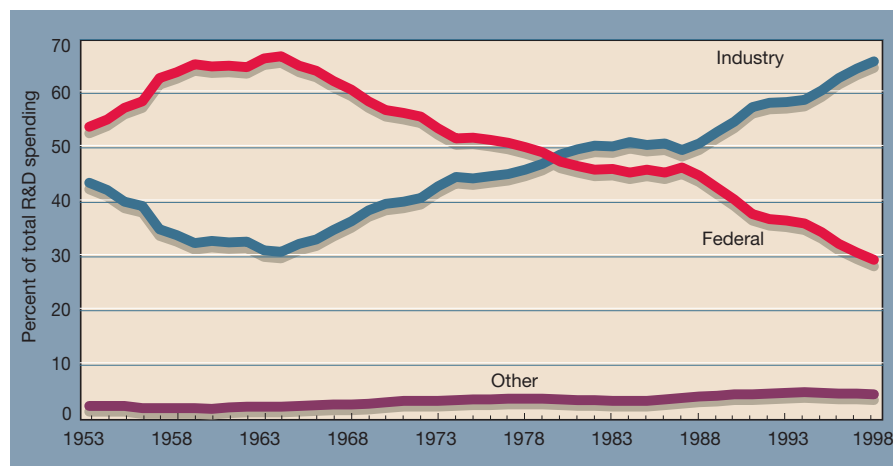


Fig 1. How industry has overtaken government in funding research in the United States.

analyses that might justify government action on environmental, health and safety issues. Should the Republicans win control of both the Congress and the presidency, one would expect them to review much of the regulatory activity not only of the Clinton administration but perhaps also of the previous George Bush (senior) administration.

Also, the Republicans can be expected to introduce into federal law limits on tort liability that have already been enacted in many states, including Texas. The two nominees' prospective judicial appointees may differ sharply in their views about what constitutes adequate scientific proof and who should be recognized as an expert in court.

It is less clear whether the two parties' differences on global climate-change policy are indicative of their overall stance on international openness and collaboration. Bush has forcefully rejected his party's opponents of immigration, whereas Gore has had to adapt only modestly to his party's opponents of economic globalization. Neither a Bush nor a Gore administration is likely to impose substantial barriers to international flows of knowledge, people, goods or money. The biggest differences may be in government-to-government cooperation. Republican Congressional concerns about security at the national laboratories, about Russian participation in the space station, and about US contributions to the Large Hadron Collider suggest that these differences might extend beyond environmental issues to military, space and even basic scientific matters.

Of course, the biggest science and technology story of the Clinton era has been the popularization of the Internet and the emergence of the 'new economy' based on it. Not surprisingly, the two parties differ substantially over whether the administration or the federal government more generally deserves credit for these developments. Gore's claim that he "took the initiative in creating the Internet" — mocked by Bush — advances, albeit clumsily, the Democratic thesis that an active federal presence in technology as well

as science is essential to the nation's economic health, even during the greatest venture-capital boom in history.

The Clinton administration's most controversial effort in this area, the Advanced Technology Program in the Department of Commerce, which matches private investments in precommercial projects, came under heavy attack in Congress, particularly from former Representative Robert Walker (Republican, Pennsylvania), who has been identified by the Bush campaign as an adviser on space and science issues. The breaking of the stalemate over this programme's budget after the 2000 election will be an accurate indicator of the larger shape of technology policy during the next administration.

Tough choices

Whoever is elected in 2000, and whomever they appoint to run key research agencies and make research and innovation policy, will face a widening range of science- and technology-intensive issues that do not divide along traditional party lines. Intellectual property rights, for instance, must be balanced between the interests of knowledge producers and those of its users, whether the product is a digital recording, genetic information or a database of telephone numbers. Privacy rights must be weighed against the demands of commerce and law enforcement. These tough choices are not eased by the application of ideological frameworks forged in the twentieth century, nor can they be comfortably deferred to the mercies of corporate decision-makers or the leisurely pace of the legal system.

The first US election of the twenty-first century might ultimately be best known as the beginning of a 'new politics' fashioned to fit the issues of the 'new economy'. Yet, for now, the shape of any such transformation is obscured by the overhang of the past century's constituencies and their baggage. ■

David M. Hart and Lewis M. Branscomb are at the John F. Kennedy School of Government, Harvard University, Cambridge, Massachusetts 02138, USA.

Supersymmetric frontiers and beyond

Towards the ultimate theory of physical laws.

Supersymmetry: Unveiling the Ultimate Laws of Nature

by Gordon Kane

Perseus: 2000. 199 pp. \$26, £15.95

Gian Francesco Giudice

The modern approach taken by physicists in attempting to understand the ultimate laws of nature has been shaped by Einstein's general relativity theory. In this, any concept of force at a distance is abandoned and symmetry emerges as a fundamental principle. Einstein had the intuition to see that a full understanding of the physical laws requires a unified treatment for gravity and electromagnetism, the two forces known at the time.

This idea of unification is still central in contemporary theoretical physics. But two new forces have since been discovered in the microscopic world: the strong force, responsible for the cohesion of nuclei, and the weak force, responsible, for instance, for neutron decay. These forces are now understood in terms of simple symmetry principles and are well described by what is called the Standard Model. Actually, the Standard Model partially achieves Einstein's dream of unification, because it identifies electromagnetic and weak forces as a single entity. As a result, the Standard Model gives a superbly successful description of particle interactions that has been verified experimentally with meticulous precision.

The Standard Model is the magnificent tool that allows today's physicists to address the most fundamental questions about the microscopic structure of matter and the evolution of the Universe, starting from a billionth of a second after the Big Bang. Nonetheless, it cannot be the ultimate theory of physical laws. As Gordon Kane states in his book, "someone who probes and studies a watch not only can describe the workings of that watch but also can say *why* the watch works". Physicists not only want to describe how the Universe works, but also want to understand why it works in that way. In other words, they want to derive the ultimate theory underlying the Standard Model, achieve a unification of all forces, and identify the fundamental symmetry principles that determine the laws of nature. As explained by Kane, the achievement of this goal may not be so far away, and supersymmetry may be a key that will lead us to it.

Supersymmetry is an unusual symmetry under which particles possessing integer spin (bosons) and half-integer spin (fermions) are exchanged. If supersymmetry holds, the laws of particle physics will be unaffected by this

special interchange of bosons and fermions. Because, among known particles, bosons are carriers of force and fermions describe matter, supersymmetry may tell us something about the relation between force and matter. In a supersymmetric world, the concept of space-time is overtaken by a new physical entity: superspace. Superspace is an extension of ordinary space-time in which new dimensions open up. These dimensions have very peculiar mathematical properties, because their coordinates are not usual numbers, but anticommuting variables (in other words, the product of two such numbers A and B satisfies the unusual property $A \times B = -B \times A$).

All this may sound very abstract. Indeed, the mathematical details of the theory are quite involved. Moreover, this is not a settled scientific theory, but a field of research in which progress is ongoing. This is, I believe, the most fascinating aspect of Kane's book. A reader with no prior knowledge of particle physics is led, always with lucid and clear explanations, from a simple introduction to the particle physics world to the frontiers of today's research. The puzzles and problems that physicists are trying to tackle are clearly revealed. Not only are readers confronted with the most recent theoretical developments in particle physics, but they can also see how physicists think in their everyday work, how they try to identify the appropriate scientific questions and proceed to find answers.

One cannot expect to learn about a scientifically established theory here. On the other hand, one can learn how basic research works and how a scientific hypothesis is developed and formulated. Of course, a scientifically sound hypothesis needs experimental scrutiny, and the theory of supersymmetry is ready for it. Kane explains how future collider experiments will be able to test the predictions made by supersymmetry. This is also a useful excuse for taking the reader on a fascinating tour of the largest accelerators in the world, in which elementary particles collide at extremely high energies and the debris products are identified and measured by sophisticated giant detectors. Most notably, the ultimate answer on the fate of supersymmetry should come from the LHC, the Large Hadron Collider under construction at CERN, the European Laboratory for Particle Physics, which will generate the highest energy collisions ever achieved.

This book gives a fascinating account of the theoretical ideas behind supersymmetry and the experimental programme aimed at its confirmation. It is told by someone who has contributed deeply to the development of the field. Kane is such an enthusiastic guide that the reader has the impression of almost participating in the research. As the tension builds, one wants to know how the story ends. But the final chapter has not yet been written. Only future research in theoretical



Filling space-time

Carpets can have many forms of symmetry, although perhaps not of the super variety. One example is shown above, from western Turkestan.

Photographs of oriental carpets of many designs and patterns are to be found in *The Carpet* by Enza Milanese (Tauris Parke/Firefly, £29.50/\$50).

book reviews

and experimental particle physics will reveal whether supersymmetry can bring us nearer to the ultimate laws of nature. ■

Gian Francesco Giudice is in the Theoretical Physics Division, CERN, CH-1211 Geneva 23, Switzerland.

The counting-house called to account

How to Build a Mind

by Igor Aleksander

Weidenfeld & Nicolson: 2000. 181 pp. £14.99

Steve Blinkhorn

Is the brain the organ of the mind? Aristotle thought it was there to cool the blood, a function it performs admirably in cold climates — witness the invention of the hat. But the common view, informed by 400 years of philosophy merging into 150 years of neurology, is that, as a settled matter of fact, the brain is the organ of the mind, with the only real puzzle being how the brain generates consciousness.

What a peculiar and uninformed position to arrive at. It suggests that people are, in a psychological sense, no more than brains on legs. The fact that adrenals, gonads and thyroid profoundly influence our conscious experience, and that sensation without sense organs is wildly discrepant and disruptive (as in the case of phantom limbs), are set aside in the hunt for the rational singularity of mind at the centre of experience.

Descartes started it, but then Locke got carried away with enthusiasm for the camera obscura, one of the original technological metaphors that supply psychologists with models when they have run out of ideas. Since then, we have been stuck with the idea that vision is the primary feedstock of consciousness, and that vision involves an internal observer making sense of what were once called 'sense data'. Anyone who has at some time lost proprioceptive and tactile sensory input over a significant part of their body knows what a loss it is, and what an important part of the sense of self it contributes. We are, as whole organisms, extended in space.

The science fiction notion of a disembodied brain floating in a nutrient solution, connected to artificial sensors by coiled cables (for some reason they are always coiled, even though the brain is immobile), is not futuristic, it is medieval. A serious examination of the functions that a disembodied immortal soul might be able to support suggests a limited range centred on incessant theological contemplation.

A disembodied brain would be little better off. Consider that, in the economy of a whole individual, the brain may be no more than the

counting-house. Accountants, the inhabitants of counting-houses, often suffer from the illusion that it is their efforts, and not those of the production-workers, sales staff and designers, that sustain the life of the enterprise. To be sure, few businesses would survive long without an effective finance function, but on its own it is utterly worthless.

Igor Aleksander has made a career out of swimming against fashionable intellectual currents so far as the notion of consciousness is concerned. When artificial intelligence was in fashion (knowledge-based systems, inference engines, 'expert systems', and the like), he was building machines on a shoestring that had emergent properties, and publishing in unfashionable places.

He works on the principle that, to make a conscious machine, you have to make a machine that is complex enough to generate its own kind of consciousness. Its desires and needs will be determined by its physical nature, not a programmer's idea of what consciousness ought to be like. The critical point comes when a machine generates internal representations of the world that are, to use Aleksander's phrase, "ego-centred": the machine represents itself to itself as a whole *vis-à-vis* a stable environment. You don't write a program to generate this style of representation: you build multiple, loosely coupled, asynchronously interacting subsystems that converge on it.

But it is appallingly difficult to capture these sorts of ideas in an accessible book and, although *How to Build a Mind* is stuffed with fine ideas, it is not a fine book. Part of the problem is the mix of methods and, in particular, the use of invented dialogues. Plato is to blame for the seductive appeal these have for authors (although, of course, he blamed Socrates); they got Galileo into awful trouble with the pope; and Aleksander isn't quite up to the literary style of Plato or the organizing genius of Galileo. Mostly, Aleksander's dialogues involve historical figures in philosophy, but they are historically unconvincing, and caricature, rather than expound, the distinctive viewpoints of their subjects. I don't know how to do it better, but it isn't done well here.

Aleksander has a point of view that stems from his discipline: engineers are people who make things that work. If he can make a machine that generates self-representations able to distinguish consistently between self and non-self, why should this not constitute a kind of consciousness? Picking up on George Kelly's dictum that "a person's psychological processes are channelised by the ways in which he anticipates events", a cat showing signs of embarrassment is surely to that extent conscious, and might not even a bee, however dimly, be representing to itself, somewhere in its 900,000 neurons, the flight path back from a food source?

Debates on mind, brain and consciousness can suffer from a lack of attention to

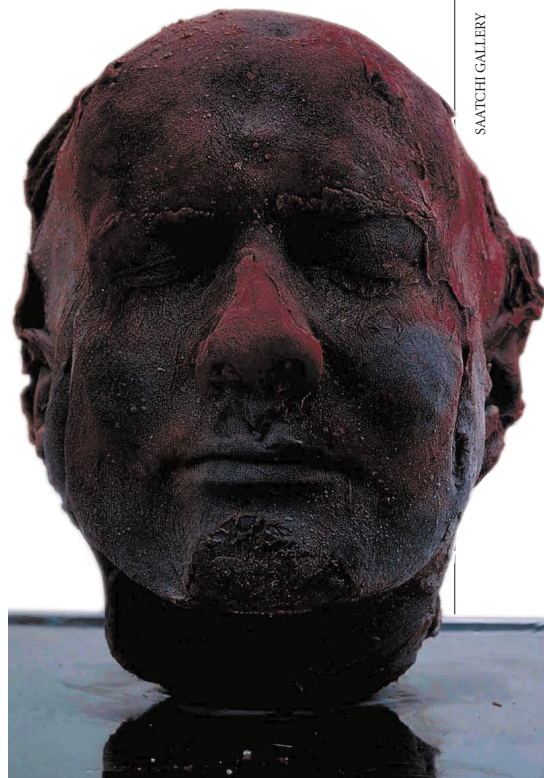
proper definition. Your notion of consciousness may differ from mine mostly in terms of our contemplation of the opposite. There are many ways of being unconscious, from the comatose to the Freudian, and we lack good distinctive terms to delineate awareness of the world, awareness of self as persistent over time and space, and the ability to exercise critical self-examination.

The overwhelming tendency in the history of philosophy, and now in the blinkered search for the means by which consciousness is generated by the brain, has been to focus on the most abstruse levels of mental functioning, such as doing mathematics, playing chess, understanding complex abstract language. But none of us got where we are without a long period of evolving consciousness and special training in a period called 'childhood', and most people spend absolutely none of their time doing any of these things anyway. Aleksander's machines have barely had the opportunity to get over their birth pangs by comparison.

What lies in the future for them is hinted at in one of the more startling of the insights that pepper *How to Build a Mind*: "if there is no perception in any sensory channel, the inner networks can fall into 'attractors' ... If this occurs during sleep, it is called dreaming."

What price a computer that dreams, or better still, that is a lucid dreamer? ■

Steve Blinkhorn is at Psychometric Research and Development Ltd, Brewmaster House, The Maltings, St Albans AL1 3HT, UK.



Sanguine contemplation: sculptor Marc Quinn's *Self*, cast in his own blood.

SAATCHI GALLERY

How the real victory went to science

The Effect of Science on the Second World War

by Guy Hartcup
Macmillan/St Martin's Press: 2000. 214 pp.
£42.50/\$59.95

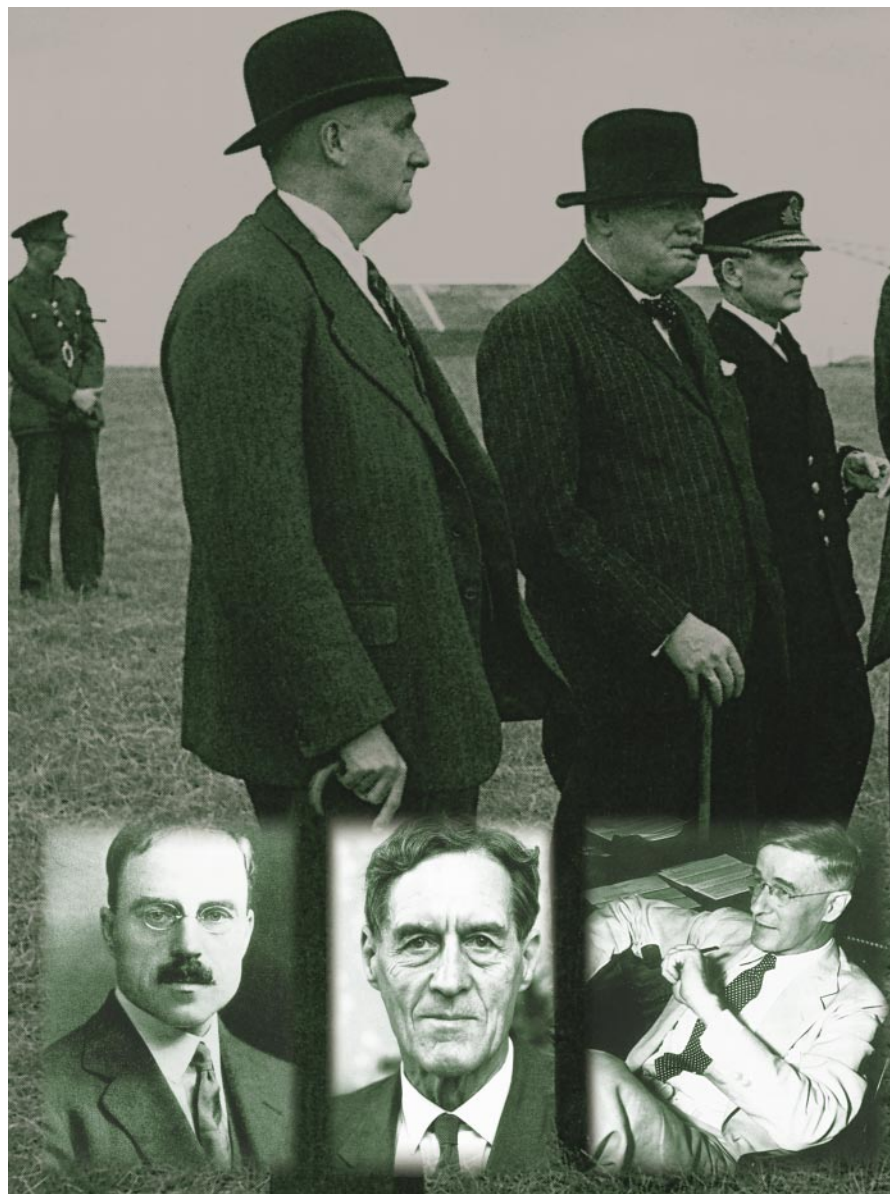
Richard Rhodes

Go to the library to find a definitive history of anything modern and you will usually come away empty-handed; academic historians, toiling in the shadows of Edward Gibbon and Thomas Macaulay, rarely risk attempting large histories any more. Guy Hartcup, an editor and government historian, finding no comprehensive assessment of the effect of science on the Second World War, decided at least to sketch one. Full credit to him for even attempting such a project in 214 pages.

Hartcup looks at organization, radio, radar, acoustic and underwater warfare, signals intelligence, operational research, medicine, chemical and biological weapons, rockets, jets and the atomic bomb, comparing Allied with Axis progress. Although breathless, occasionally muddled and disappointingly inconclusive, the result is a useful compendium of the remarkable variety of contributions scientists made between 1930 and 1945 to both defensive and offensive war.

"Success," Hartcup argues, "went to the side which recognized the importance of close collaboration between scientists and the military, something unknown before 1939." Certainly, the German and Japanese military leadership arrogantly resisted scientific collaboration. But scientists eager to contribute encountered at least incomprehension even in Britain and the United States. British radar development, Hartcup reports, began with a search for a death ray. The challenge was fortunately directed to R. A. Watson-Watt, who had ideas for using electromagnetic radiation in air defence. In the United States, death rays were also evoked when a US Army ordnance officer on a committee set up to consider atomic bomb research disparaged the expert judgement of young Leo Szilard, Eugene Wigner and Edward Teller. The three émigré physicists, worried about German fission research, were seeking modest funding for work towards a slow-neutron chain reaction that was being carried out by Enrico Fermi at Columbia University. The officer countered that he had a tethered goat and a big prize waiting for anyone who could kill the goat with a death ray. "Nobody has claimed the prize yet," he smirked. Such scepticism about science's value in war delayed Fermi's funding for months.

A more insidious military resistance was the fanatical devotion of the Allied air forces



Science at war: some leading protagonists — top, Winston Churchill with his science adviser, F. A. Lindemann; below, left to right, Henry Tizard, P. M. S. Blackett and Vannevar Bush.

to strategic bombing at the expense of defence and tactical support. The priority given to centimetric radar as a strategic bombing aid meant that the installation of centimetric shore-based radars that could have significantly reduced German U-boat depredation on Allied shipping was delayed by at least six months.

A year later, during the run-up to the Normandy invasion, when operations research demonstrated the vulnerability of French railroad systems to the bombing of railroad maintenance centres, strategic bombing commanders refused to release aircraft to attack them because they believed their bombing campaign would soon win the war. Fortunately, Air Chief Marshal Sir Arthur Tedder and Supreme Allied Commander Dwight Eisenhower intervened and insisted on bomber diversion to the railroad mission. This resulted in a 20% reduction in

the total volume of French rail traffic, according to Hartcup, and delayed the arrival of German tank and infantry divisions on the Normandy peninsula.

Hartcup is almost comically chauvinistic in assessing the weight and priority of British and US scientific contributions. Certainly Britain, fighting for her life early in the war while the United States remained unmobilized and officially neutral, usually led the way. The British invention of the cavity magnetron, fundamental to microwave radar, and the willingness to share it resulted in the founding of the Radiation Laboratory at MIT, which became a cornucopia of life-saving inventions.

The early, accurate, theoretical assessment of the prospects for an atomic bomb by the émigré physicists Rudolf Peierls and Otto Frisch at the University of Birmingham, which their colleague Mark Oliphant passed

directly to Henry Tizard, the chief scientific adviser to the defence ministry, led to the famous 1941 MAUD Report, which laid out the path to the development of the bomb. It was this that finally convinced Franklin Roosevelt and his advisers to authorize what came to be called the Manhattan Project.

But crediting the University of Cambridge's Cavendish Laboratory with discovering plutonium confuses speculation with investigation. And asserting that the man-made element was "also" discovered by Glenn (not 'Glen') Seaborg is simply wrong. Seaborg and his colleagues at the University of California at Berkeley were the first to precipitate a measurable quantity of element 94 from thorium, the accepted marker of discovery, on 5 March 1941. Both Carl Friedrich von Weizsäcker in Germany and the Cavendish team independently realized early on that transuranics would be highly fissionable, but the MAUD Report focused on isotope separation and a uranium bomb, and that was what Oliphant pushed when he came to the United States in August 1941. Plutonium production was authorized in early December 1941 only because the US physicists Ernest Lawrence and Arthur Compton vouched for it.

Hartcup's implicit model of Allied co-operation is a canny British head attached to a headless but muscular American body. In fact, inventions and discoveries emanated from both sides of the Atlantic as the two nations cooperated at unprecedented levels of trust.

Radar, pioneered in Britain, had the greatest effect on the war. "Radar won the war," US scientists say, "but the atomic bomb ended it." Surely there is more than enough credit to go around.

I found too many mistakes and misstatements in this brief book. Alan Turing's proposal for a punched-tape logic machine did more than merely "anticipate" the modern computer: it worked out its fundamental logic. The scarcity of uranium-235 in natural uranium is not the reason early theoretical estimates of bomb size were so large: people were still thinking in terms of slow neutron fission, which would have required weaponizing a nuclear reactor. (The crucial message Oliphant delivered on his mission to the United States was fast fission with pure uranium-235, a sphere of which has a tamped critical mass of only 15 kilograms.) Heavy water and other moderators don't "limit the multiplication of neutrons", but decelerate and reflect them.

Also, the implosion method of detonating a bomb was not "the responsibility of the British team" at the Los Alamos Laboratory, and although Peierls and Klaus Fuchs contributed to implosion theory there, it was not they alone who "made the implosion possible". Laboratory head Robert Oppenheimer redirected almost the entire Los Alamos staff

Science in culture

Realities in wax

A comparative anatomical exhibition at the Deutsches Museum Bonn.

Alison Abbott

When La Specola museum opened its doors to the Florentine public for the first time in 1775, visitors were confronted with what would have probably been their first view of the inside of the human body. They saw detailed, anatomically correct, wax models of body parts, fed by their networks of blood vessels and supported by bones, tendons and muscles.

The collection was commissioned in 1771 by Grand Duke Leopold I of Tuscany, and over the next decades grew to more than 1,400 specimens. It was housed, along with other scientific exhibits, in the grand duke's *Wunderkammer*: public access to the new scientific knowledge was in the spirit of the times, which became known as the Age of Enlightenment.

Returning from Florence in 1786, Goethe reported that "three-dimensional anatomy ... has been practised in Florence for many years at a very high level, but it can only flourish where science, art, taste and technology are integrated in living practice".

The exhibition at the Deutsches Museum Bonn is part of the 'new' movement to bring together the worlds of art and science. The volume of knowledge is now so vast that the interface between the two is not as self-evident as that which sparked Goethe's words of admiration. But the exhibition nevertheless makes a salient point with its juxtaposition of 30 samples of the Florentine wax models — which have never before been permitted to leave their home town — with modern images of the body as seen from within: two-dimensional X-ray photos (some from 1895), three-dimensional computer magnetic resonance tomography,

positron-emission tomography, angiography and the like.

Those naive eighteenth-century visitors would have been more impressed by their first confrontation with internal anatomy than we, with our overexposed and dulled senses, could ever be by more sophisticated, and theoretically more spectacular, medical images.

Alison Abbott is the European correspondent of Nature.

"La Specola: Anatomie in Wachs im Kontrast zu Bildern der modernen Medizin" (Anatomy in Wax in Contrast with Images of Modern Medicine) runs until 19 November 2000.



DEUTSCHES MUSEUM BONN

to implosion work in the summer of 1944; the problem was that difficult. Hartcup properly blames Fuchs for passing secrets to the Soviet Union, but omits to mention the fully equivalent treachery of the American Theodore Hall.

Enough errors of this kind mar the atomic-bomb chapter of Hartcup's book to leave me wondering how reliable he is on subjects I know less about. Gas and bacteria Hartcup sees as "unacceptable weapons". Would that they had been. Applied narrowly to combat, that assessment is accurate, but it should not be forgotten that gas — pure carbon monoxide, carbon monoxide from engine exhaust, and cyanogen-releasing Zyklon — was used to murder millions of civilians in the gas vans and gas chambers of the Third Reich.

Hartcup's three-page conclusion, which purports to include the war's "aftermath", is

far too modest. Science did much more than help win the war. By inventing the nuclear reactor, it made available the first new and practically unlimited energy source in human history. By inventing the atomic bomb (and the thermonuclear bomb that followed), it established a natural threshold beyond which war would be suicidal.

Harnessing science to war, the nation-states sought to enhance their power. Instead, science challenged the nation-state, simultaneously limiting its sovereignty and offering its populations the prospect of an abundance that may eventually remove the economic inequities that are the fundamental cause of war. The Second World War made science the most powerful political institution humankind has yet devised.

Richard Rhodes is at 609 Summer Hill Road, Madison, Connecticut 06443, USA.

Magic beans

The Japanese invader that's good for you.

Vaclav Smil

In July 1853, Commodore Matthew Perry's small flotilla cast its anchors off Uraga at the entrance to Edo (Tokyo) Bay. Its mission was to end Japan's quarter of a millennium of isolation under the Tokugawa shogunate. The black ships returned the next spring, when Perry signed the treaty. The shogunate was persuaded not just by the ships' large guns but also by the demonstrations of American technical superiority that included the telegraph and a miniature steam locomotive.

Mission accomplished, Perry returned bearing samples of Japanese products. As well as the predictable silks, lacquer boxes and exquisite porcelain, these included white and red soy beans. Several months later the US Patent Office was offering the seeds to American farmers, but there were few takers. Three decades later, Lewis Sturtevant noted in his exhaustive survey of edible plants that soy beans had begun to appear as a novelty item in US seed catalogues.

In China their cultivation is ancient: nearly 5,000 years ago the legendary Emperor Shennong made soy beans the only legume of his five life-sustaining grains, and millennia of culinary evolution have produced an enormous range of soy-based foodstuffs throughout East and Southeast Asia. Immature pods are eaten as a vegetable. Mature seeds are sprouted or milled to make flour. Protein in milled soy beans can be precipitated by coagulants, usually calcium and magnesium sulphate, to make bean curd — Japanese tofu and Chinese *doufu*. Seeds are ground or boiled and then fermented to prepare soy sauces and pastes such as Japanese miso and natto and Indonesian tempeh. And the seeds, which are nearly 20% lipids, make an excellent cooking oil consisting of 85% unsaturated fatty acids. Modern processing has also added an array of flavoured drinks to traditional soy bean milk.

The West was slow to adopt soy beans: in the 1920s American farmers still devoted more land to sweet potatoes. Soy beans finally took off in the 1930s, but it was the prosperity following the Second World War, with its rising demand for cheaper meat, that made them a major crop. Corn has only 10% protein, but most soy bean cultivars have about 36%, nearly twice as much as other legumes. Mixtures of corn, a high-yielding source of carbohydrates, and highly proteinaceous (44%) soy bean meal, produced by first extracting oil with organic solvents, have come to dominate feeding formulas

around the world. American cattle, German broilers, Japanese pigs — and Chinese carp — all yield more meat thanks to *Glycine max*.

The US soy bean harvest passed 10 million tonnes (Mt) in the early 1950s. It is now 70–75 Mt a year — almost half of the global total. Soy beans are the country's second most valuable crop, close behind corn and worth nearly three times as much as wheat, and grow on more than 15% of all arable land. Brazilian production, geared to exports, has grown even faster. The area planted with soy beans has expanded more than 60-fold since the late 1950s; Brazil now produces about 30 Mt. Argentina harvests about half as much as this, relegating the Chinese harvest to fourth place. The Japanese now grow a mere 3% of what they consume. Boiled tofu served in the temple precincts of Kyoto's Higashiyama or tofu-based nouvelle cuisine in Tokyo's Ginza now come from seeds harvested, and often organically grown, in Iowa or North Carolina.

The symbiotic relationship between soy beans and the *Rhizobium* bacteria in nodules

attached to the plant's roots can provide as much as 100 kg of nitrogen per hectare per year. So it may come as a surprise to learn that at least a fifth of the US crop now regularly receives nitrogen fertilizers. This is because a higher proportion of the biomass of modern cultivars is in the seed, which is removed from the field. So, even with minimal losses to leaching or erosion, more nitrogen is extracted than is fixed by bacteria. In contrast, Brazilian soy beans receive no extra nitrogen. Perhaps the most important environmental downside of soy beans' rise is that, as with all other crops grown in rows, their cultivation can lead to higher soil erosion.

The world's diet is unthinkable without soy beans, but they remain a fringe foodstuff in the West. Only soy sauce has become a universal condiment — bean curd, although readily available, is still a culinary oddity. A change would serve women particularly well. Soy bean foods contain very high concentrations of isoflavones, particularly genistein. These phytoestrogens can alleviate the symptoms of menopause, and their consumption has been linked to reduced risks of breast cancer in East Asia. Soy proteins also lower levels of blood cholesterol, and some, although not all, epidemiological studies link their intake with a reduced risk of cardiovascular diseases. Greater enthusiasm for the subtle taste of tofu could thus have highly cost-effective health benefits.

Vaclav Smil is at the University of Manitoba, Winnipeg, Manitoba R3T 2N2, Canada.

Soy beans are the second most valuable US crop, and grow on more than 15% of all arable land.



Soy ahoy: the beans reached the West via Perry's 'black ships', shown here arriving off Tokyo in 1853.

Programmable matter

How quantum wellstone ushered in our modern age.

Wil McCarthy

4 July 2100. The flick of a switch: a wall becomes a window becomes a door. Any chair becomes a hypercomputer, any rooftop a power or waste-treatment plant. We scarcely notice; programmable matter pervades our homes, our workplaces, our vehicles and environments. There isn't a city on Earth — or Mars, for that matter — that isn't clothed in the stuff from rooftop to sub-basement. But although we rarely stop to consider it, the bones of these cities — their streets, their sewers, the hearts of their telecom networks — were laid out in a time when the properties of matter were dictated exclusively by mother nature.

Imagine: if specific mechanical or electrical properties were desired, one first had to hire miners to extract appropriate elements from the Earth, then chemists and metallurgists to mix precise proportions under precise conditions, then artisans to craft the resulting materials into components, and assemble the components into products that could then be transported to the location of desired use. The inconvenience must have been staggering.

In the 22nd century, of course, any competent designer will simply define the shape and properties they require, including 'unnatural' traits such as superreflectance, refraction matching (invisibility), and centuple bonds far stronger than diamond, then distribute the configuration file to any interested users. But prior to the invention of wellstone — the earliest form of programmable matter — this would have been pure fantasy. With this in mind, we'll look back on the invention upon which, arguably, our entire civilization rests.

Consider silicon, whose oxide is the most abundant material on Earth. Humans had been making hammers and millstones out of it for millions of years, but, as it happens, silicon is also a semiconductor. The electrical properties of pure silicon are fixed by the laws of physics, but through doping — the carefully controlled introduction of impurities — its crystals can be tuned so that, for example, room-temperature electrons have a good chance of jumping up into the conduction band.

Although invaluable in the development of twentieth-century digital electronics, silicon's killer application eventually proved to be as a storage medium for electrons. The layering of doped silica in particular ways

can trap conduction electrons in a membrane so thin that, from one face to the other, their behaviour as tiny quantum wave packets takes precedence over their behaviour as particles. This is called a 'quantum well'. From there, confining the electrons along a second dimension produces a 'quantum wire', and finally, with three dimensions, a 'quantum dot'.

The unique trait of a quantum dot, as opposed to any other electronic component, is that the electrons trapped in it will arrange themselves as though they were part of an atom, even though there's no atomic nucleus for them to surround. Which atom they emulate depends on the number of electrons and the exact geometry of the wells that confine them, and in fact where a normal atom is spherical, such 'designer atoms' can be turned into cubes, tetrahedra or any other shape, and filled with vastly more electrons than any real nucleus could support, to produce 'atoms' with properties that simply don't occur in nature.

Significantly, the quantum dots needn't be part of the physical structure of the semiconductor; they can be maintained just above it through a careful balancing of electrical charges. In fact, this is the preferred method, since it permits the dots' characteristics to be adjusted without any physical modification of the substrate.

Who invented wellstone remains a matter of debate: similar work was being performed in parallel, in laboratories all over the world. Regardless of whose actual idea it was, the concept itself is deceptively simple: a diffuse lattice of crystalline silicon, superfine

threads much thinner than a human hair, criss-crossing to form a translucent structure with roughly the density of balsa wood. Like balsa wood, the structure consists mostly of empty space, except that with the application of electrical currents, that space can be filled with 'atoms' of any desired species, producing a virtual substance with the mass of diffuse silicon, but the chemical, physical and electrical properties of some new, hybrid material.

Wellstone iron is weaker than its natural counterpart, less conductive and ferromagnetic — essentially, less 'iron-like' — and if you bash it over and over with a golf club it will gradually lose any resemblance to iron, reverting to shattered silicon and empty space. On the other hand, it's feather-light, wholly rustproof, and changeable at the flick of a bit into zinc, rubidium, or even 'imaginary' substances like impervium, the toughest superreflector known.

Of the changes wrought by programmable matter in the past century, not all have been universally welcomed. Wellstone, in the grand Promethean tradition, places the power of creation and destruction squarely into human hands. Many have argued that, far from making us strong, this power fosters a quiet corruption of spirit. Still, the fable of the three little pigs holds true — not even the Luddites among us build their houses of straw or sticks, when impervium is a free download. ■

Wil McCarthy, an aerospace engineer and science columnist, is the author of six science fiction novels, including Bloom (Millennium) and The Collapsium (Del Rey/Gollancz).



Uncertainty in climate change

Andrew J. Weaver and Francis W. Zwiers

Climate modelling has produced varying projections of possible global rises in temperature. A simple statistical method brings several such projections into closer agreement and includes an indication of potential accuracy.

Governments around the world are investing heavily in coupled-climate models to project future climate change. Such models have interacting atmosphere, ocean, land and sea-ice components, and serve as laboratories for studying the effects of natural and human influences on the climate system. Not surprisingly, however, different models produce different results. This diversity is useful scientifically, because it challenges our understanding of the physical processes that govern climate. But what are policy-makers to make of the different projections? Is one of them right? Are all of them wrong? In short, can uncertainty estimates be put on such projections? Previous methods to estimate uncertainty, which have used the subjective opinions of experts, simplified climate models or the spread between coupled-climate model projections, are all less than satisfactory.

Climate models have to take into account numerous physical processes, including the radiative processes that transfer energy through the atmosphere. Greenhouse gases, such as CO₂ and chlorofluorocarbons, and particulates, such as sulphate aerosols, all affect this energy transfer. An atmosphere that is enriched in greenhouse gases may lose less heat to space and consequently become warmer. Aerosols may either warm or cool the atmosphere, depending on their properties. Sulphate aerosols have a direct cooling effect by scattering incoming sunlight back to space. They may also have an indirect cooling effect by influencing the lifetime and reflectivity of clouds. Natural 'forcing' factors, such as volcanic eruptions and variations in solar output, may similarly influence climate.

On page 617 of this issue, Allen *et al.*¹ provide a simple way of estimating the uncertainty in projections of twenty-first-century climate. This powerful use of statistics brings estimates from several modelling groups into line. It also leaves us with a projected global warming, relative to preindustrial times, of 1–2.5 °C in the decade 2036–2046, with an uncertainty range of 5–95%. That is, the estimation procedure is designed to produce warming-range estimates that will contain the true warming nine times out of ten in repeated, independent analyses.

Allen *et al.*¹ start with the observed and modelled near-surface air temperatures for each decade in the period 1946–96 and

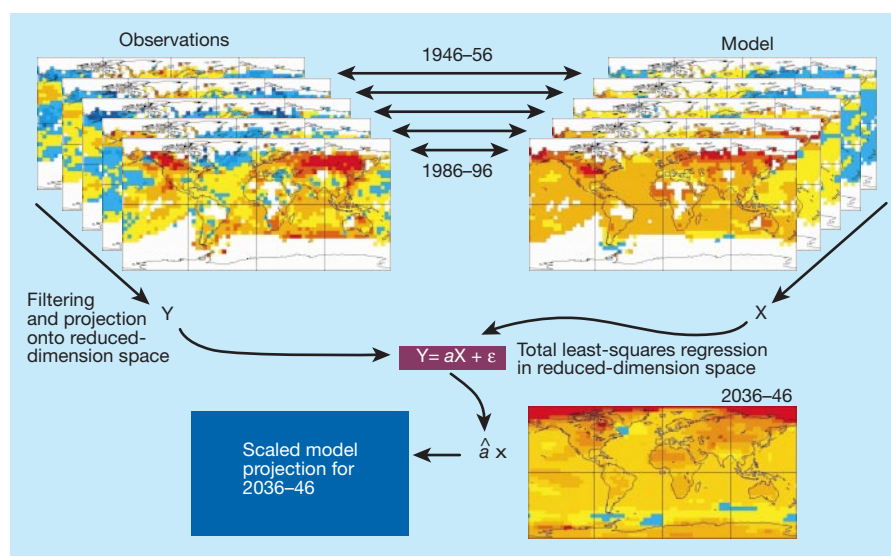


Figure 1 The approach of Allen *et al.*¹ to estimating temperature for 2036–2046, and the uncertainty of that estimate. They first find the differences between decadal averaged temperatures starting with 1946–1956 and the century average (upper left). A model incorporating the effects of observed changes in the concentrations of greenhouse gases and sulphate aerosols is then run to simulate the climate response to these factors (upper right). These maps are filtered and reduced to a form that retains the essential observed (Y) and simulated (X) information on climate change over the period 1946–96. Regression analysis is used to estimate the scaling factor a that produces the best match between observations and the simulated climate-change signal. The discrepancy ϵ between the observations and the scaled simulation results from natural variability in the climate system and model errors. Long 'control' simulations (in which greenhouse gas concentrations are held constant) are used to assess the uncertainty of the scaling factor. Finally, model projections of a future decade can then be scaled by the estimated factor ($\hat{a}$), and its uncertainty used to express the uncertainty of the scaled projection.

express them as anomalies relative to the twentieth-century mean. Their strategy is then to find out what constant scaling factor of model-simulated temperatures is required to maximize the agreement between the observed and simulated decadal mean temperatures for 1946–96 (Fig. 1). The multiple regression used to determine this model-specific scaling factor also yields an uncertainty estimate. Allen and colleagues then show that in a simplified climate model there is a simple linear relationship between the size of the twentieth-century change and that projected for 2036–46.

The beauty of this result is that it is independent of the 'climate sensitivity' of the model (defined as the eventual warming that occurs as a result of doubling the concentration of atmospheric CO₂) and the rate at which that warming happens. In particular, it is independent of the rate of oceanic heat uptake, which varies between models with different representations of ocean physics.

More complicated climate models have indeed exhibited a linear response to changes in radiative forcing caused by increases in atmospheric greenhouse gases and sulphate aerosols. For example, when separate model runs with twentieth-century levels of greenhouse gases and aerosols are combined, the result is similar to runs which included them together².

Early model simulations driven exclusively by greenhouse gases tended to overestimate the warming in the twentieth-century record. In 1995, two studies appeared^{3,4} which showed better agreement when the direct effects of sulphate aerosols were also included. Since then some modelling groups have also included the effects of sulphate aerosols on cloud properties, and many now represent greenhouse gases individually, rather than using CO₂ as a surrogate for all anthropogenic greenhouse gases.

Modellers are now also taking into account the historical variation in natural

influences on climate. For instance, some studies^{5,6} suggest that changes in solar irradiance may have been a factor in the warming in the early part of the century. Aerosols injected into the upper atmosphere by volcanic eruptions, and changes in ozone, are also components that need to be included to reproduce some aspects of variability in the climate system⁷.

This approach of adding in additional factors to reproduce the twentieth-century record has been criticized as being merely a curve-fitting exercise. However, Allen and colleagues' projected temperature change for the middle of the twenty-first century appears to be insensitive to whether or not a forcing or process is missing (such as the effect of sulphate aerosols on clouds, or of variations in solar or volcanic influences). At the least, they argue, this should be the case provided that the climate feedbacks of the missing process or forcing are close to linear for small departures from the present climate. Of course, all these conclusions fall apart if a sudden, yet highly improbable, nonlinear transition between climatic regimes occurs. Such a transition might occur, for instance, if the North Atlantic 'conveyor' were to shut down. The conveyor transports warm surface water northwards and cold deep water southwards, and has a large influence on climate.

In 1996, the United Nations Intergovernmental Panel on Climate Change (IPCC) published a report⁸ in which the controversial statement "The balance of evidence suggests a discernible human influence on global climate" appeared. Many studies (for instance refs 5 and 6) have since provided strong support for this view. But policy-makers care little about the intricacies of scientific modelling and data analysis. Rather, they require future climate projections to come with quantitative uncertainty estimates. Allen *et al.* have provided a way forward to that end.

Climate-model projections, of course, are only as good as the 'scenarios' used to infer future levels of anthropogenic greenhouse gases and aerosols. The study by Allen *et al.* relies on the so-called 'business-as-usual' scenario set forth in 1992 by the IPCC⁹, in which greenhouse gas concentrations grow at about 1% compounded annually, and aerosol concentrations in industrialized areas increase substantially until the middle of this century. Earlier this year, however, a series of 40 different emission scenarios were published¹⁰, based on a variety of forecasts of socioeconomic change, energy use, population growth and technological advance. It will be impossible for modellers to undertake the necessary ensemble of integrations for the full range of scenarios, so we can expect to see only a few illustrative examples emerging in the years to come. But we can now hope to have

quantitative uncertainty estimates attached to them.

Andrew J. Weaver is at the School of Earth and Ocean Sciences, University of Victoria, PO Box 3055, Victoria, British Columbia, V8W 3P6, Canada.

e-mail: weaver@uvic.ca

Francis W. Zwiers is at the Canadian Centre for Climate Modelling and Analysis, Meteorological Service of Canada, c/o University of Victoria, PO Box 1700, Victoria, British Columbia, V8W 2Y2, Canada.

1. Allen, M. R., Stott, P. A., Mitchell, J. F. B., Schnur, R. & Delworth, T. L. *Nature* **407**, 617–620 (2000).
2. Haywood, J. M., Stouffer, R. J., Wetherald, R. T., Manabe, S. & Ramaswamy, V. *Geophys. Res. Lett.* **24**, 1335–1338 (1997).

3. Mitchell, J. F. B., Johns, T. J., Gregory, J. M. & Tett, S. B. F. *Nature* **376**, 501–504 (1995).
4. Hasselmann, K. *et al.* in *Proceedings of "Modern Dynamical Meteorology", Symposium in Honor of Aksel Wiin-Nielsen, 1995* (ed. Ditlevsen, P.) 201–221 (European Centre for Medium-range Weather Forecasts, Reading, UK, 1995).
5. Tett, S. F. B., Stott, P. A., Allen, M. R., Ingram, W. J. & Mitchell, J. F. B. *Nature* **399**, 569–572 (1999).
6. Hegerl, G. C. *et al.* *Clim. Dyn.* **13**, 613–634 (1997).
7. Hansen, J., Sato, M., Lacis, A. & Ruedy, R. *Phil. Trans. R. Soc. Lond. B* **352**, 231–240 (1997).
8. Intergovernmental Panel on Climate Change *Climate Change 1995: The Science of Climate Change* (Cambridge Univ. Press, 1996).
9. Intergovernmental Panel on Climate Change *The Supplementary Report to the IPCC Scientific Assessment* (Cambridge Univ. Press, 1992).
10. Intergovernmental Panel on Climate Change *Special Report on Emission Scenarios* (Cambridge Univ. Press, 2000).

Cognitive science

The logic of human learning

Nick Chater

We view the world not as a continuous flux, but in terms of discrete, labelled categories, such as tables, trees and teapots. These categories are by no means arbitrary — a 'pseudo'-category consisting of an arbitrary assortment of objects starting with the letter 't' would not only be extremely difficult to learn, it would be hard to remember and to interpret in new situations. But what makes some categories 'better' than others? Why are we disposed to divide the world into teapots and not 't-words'? On page 630 of this issue, Jacob Feldman¹ takes an important step towards answering this question. He suggests that there is a formal measure of complexity that determines how natural a category is and how difficult it is to learn.

Feldman is specifically concerned with

the important, although limited, class of 'boolean concepts' — that is, categories that can be defined in terms of logical operations such as AND, OR and NOT. In the experiment, subjects were presented with images of 'amoebas', which differed in binary features, such as shape of the nuclei, number of the nuclei and so on (Fig. 1). Boolean categories were then defined in terms of these features. So, with features a, b and c we might define the boolean concept (a AND b) OR (NOT c). This concept applies to items that either have properties a and b, or do not have property c, or both. This concept can, of course, be logically expressed in (infinitely) many ways — for example, as NOT(c AND ((NOT a) OR (NOT b))).

Feldman pursues the intuitively attractive idea that simple boolean concepts are the easiest to learn. But how can complexity be measured? First, the complexity of a logical formula is defined as the number of non-logical terms (the primitive features a, b and c) that occur in that formula — this measure is boolean complexity. So, for example, the formula (a AND b) OR (a AND (NOT b)) would have a boolean complexity of 4. Second, we can define the complexity of a boolean concept as the minimal complexity of any logical formula that expresses that concept. The boolean concept associated with (a AND b) OR (a AND (NOT b)) can most simply be expressed by a very simple formula: a. Hence it has a boolean complexity of 1. Feldman presents impressive experimental data showing that the greater the boolean complexity of a concept, the harder it is to learn. So it seems that boolean complexity provides a measure of the 'naturalness' of a concept.

Feldman has established an important result in the restricted domain of boolean concepts. How far might this result general-

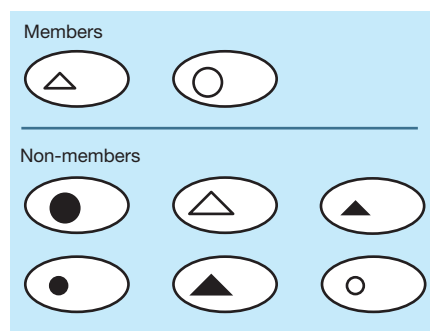


Figure 1 Learning complex rules. In Feldman's experiments¹, subjects were presented with images of 'amoebas', which were defined by simple binary features (such as the shape and size of nuclei). Subjects had to learn to distinguish members of a particular 'species' of amoeba from non-members. Each species corresponded to a 'boolean category' of varying logical complexity. Feldman shows that the difficulty of learning a particular concept is related to its logical complexity.

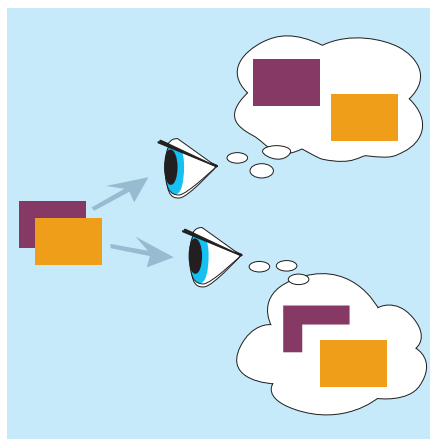


Figure 2 Simplicity in perception. The visual stimulus on the left can be interpreted as one square tile 'on top' of another or as an 'r' shape abutting a tile. The preference for the 'two tile' interpretation may be due to its simplicity — two squares are simpler to specify than an 'r' and a square. Feldman¹ suggests that this kind of simplicity principle, already widely applied in perception, may also explain how people learn to divide the world into discrete categories.

ize to learning other kinds of concepts, or to learning across the board? Some considerations suggest caution. First, boolean concepts are very rare among natural language concepts. We simply cannot define, say, a face as a boolean combination of mouth, nose or other features, despite vast efforts to provide such definitions in philosophical analysis and artificial intelligence². Second, real-world concept learning may draw on innate content-specific knowledge, such as the knowledge that biological categories can be usefully organized into nested hierarchies (for example, from dachshund to dog to animal to living thing)³.

More persuasive, though, are indications that Feldman's result may form part of a very general theory of cognition. Feldman notes that boolean complexity is closely related to the more general idea of Kolmogorov (or K) complexity⁴. The K-complexity of an object is the shortest computer program that can generate that object. So, Hamlet would have high K-complexity, as there is no short program that can generate it; on the other hand, a sequence of a billion zeroes has low K-complexity, because it can be produced by a very short computer program.

If the 'programming language' is elementary logic, and we assume that the objects to be encoded are boolean concepts and that the code length is dominated by the non-logical terms, then K-complexity approximates to boolean complexity. But K-complexity is a more general notion that applies not just to boolean concepts, but to representations of any kind: logical, linguistic, probabilistic or pictorial. As a purely abstract theory, K-complexity has led to methods for inductive inference, based on

the search for the simplest interpretation of the current data, which are theoretically well justified and effective in machine learning and statistics^{5,6}.

Feldman's results may therefore be a special case of a general principle: that human cognition favours simple interpretations of the world, or interpretations involving low K-complexity⁷. This view has been widely advocated in the psychology of perception^{8,9}. For example, simplicity may determine the preferred interpretation of ambiguous figures (Fig. 2). In language acquisition, it has been proposed that children search for patterns or regularities that 'compress' their linguistic input¹⁰. And in physiology it has been argued that finding short codes for

sensory input may be a fundamental goal of the brain¹¹. Feldman's work may therefore relate not only to categorization, but also to fundamental principles of human learning. ■

Nick Chater is in the Department of Psychology, University of Warwick, Coventry CV4 7AL, UK.

e-mail: nick.chater@warwick.ac.uk

1. Feldman, J. *Nature* **407**, 630–633 (2000).
2. Fodor, J. A. *Concepts* (Oxford Univ. Press, 1998).
3. Atran, S. *Behav. Brain. Sci.* **21**, 547–609 (1998).
4. Li, M. & Vitanyi, P. *Introduction to Kolmogorov Complexity and its Applications* (Springer, Berlin, 1997).
5. Quinlan, J. & Rivest, R. *Inform. Comput.* **80**, 227–248 (1989).
6. Rissanen, J. *IEEE Trans. Inform. Theory* **42**, 40–47 (1996).
7. Chater, N. Q. *J. Exp. Psychol.* **52A**, 273–302 (1999).
8. Leeuwenberg, E. & Boselie, F. *Psychol. Rev.* **95**, 485–491 (1988).
9. Chater, N. *Psychol. Rev.* **103**, 566–581 (1996).
10. Brent, M. R. & Cartwright, T. A. *Cognition* **61**, 93–126 (1996).
11. Barlow, H. B., Kaushal, T. P. & Mitchison, G. J. *Neural Computat.* **1**, 412–423 (1989).

Meteorites

The long trip to Earth

Clark R. Chapman

In the 1790s, European scientists accepted the once preposterous notion that meteorites — rocks that fall from the skies — were of cosmic origin and not lightning bolts, volcanic ejecta, or the eighteenth-century equivalents of ghosts or UFOs. By coincidence, the first asteroid was discovered in space in 1801. As more asteroids were found, the idea naturally developed that meteorites were fragments of asteroids. Early theories that a planet had exploded to create the asteroid belt between Mars and Jupiter gave way to the modern understanding that asteroids are the left-over building blocks of a planet that never formed. Although meteoriticists continued to assume that these exotic rocks were asteroidal, physicists increasingly doubted that rocks could be moved intact from the asteroid belt to Earth. After 200 years of disparate astronomical observations of asteroids and laboratory studies of meteorites, a quantitative model, reported by David Vokrouhlický and Paolo Farinella on page 606 of this issue¹, now shows how a suite of processes bring small asteroid fragments to Earth.

From a physics perspective, the simple billiard-ball analogy of collisions between rocks knocking each other around the inner Solar System does not stand up to elementary scrutiny. It is no easier to 'bump' icy, rocky or even metallic objects, with finite material strengths, from the asteroid belt into Earth-crossing orbits than it is to hit eggs around the fairways with a golf club. Much gentler accelerations are required, such as those exerted on comets at the edge of the Solar System by the gravity of passing stars and giant molecular clouds.

By the 1970s physicists had realized that periodic gravitational forces produced by

Jupiter and Saturn could gently move an asteroid fragment located in one of two specific narrow zones (known as 'orbital resonances') into an elongated orbit that might intersect with that of the Earth. At the very least, the fragment might pass near Mars, which in turn could gravitationally deflect it into an Earth-crossing orbit. In 1985 Jack Wisdom² used chaos theory to show how such transfers might work. Indeed, nowadays, asteroids are missing from these resonant 'escape hatches', providing evidence that asteroids were cleared out this way in the past.

The picture that emerged showed how collisions involving asteroids such as 6 Hebe, located at the edge of a resonance, would eject fragments at tens to hundreds of metres per second (consistent with the degree of shock seen in meteorites) into the resonance gap for quick delivery to Earth. If true, then meteorites must represent just a few, favourably located asteroids, not the vast majority, which are far from the resonances. But the Earth encounters a surprisingly large amount of meteoritic material, perhaps 1,000 tons per year. For this to come from just a few asteroids requires an unnaturally high transfer efficiency and model parameters had to be tweaked uncomfortably to make the numbers work. The selectivity of the resonance process at least helped to explain why astronomers could not find main-belt asteroids with the same spectral attributes as the most common meteorites in our museums, the ordinary chondrites. But it undercut meteoriticists, whose theories assumed that ordinary chondrites were common asteroidal materials.

Ten years ago, the advent of fast computers gave new impetus to this research, but a problem soon emerged. Farinella and



100 YEARS AGO

Prof. John Perry has asked me to write something in criticism of the views he has lately expressed about the teaching of mathematics... It is shocking that young people should be adding their brains over mere logical subtleties, trying to understand the proof of one obvious fact in terms of something equally, or, it may be, not quite so obvious, and conceiving a profound dislike for mathematics, when they might be learning geometry, a most important fundamental subject. I hold the view that it is essentially an experimental science, like any other, and should be taught observationally, descriptively and experimentally in the first place... The value of π should be measured; it may be done to a high degree of accuracy. So with the area of the circle, ellipse and all sorts of other things... The boy who really measures and finds it true will have grasped the fact far better than by a logical demonstration without adequate experimental knowledge; for it happens that boys, who are generally very stupid in abstract ideas, learn a demonstration without knowing what it is all about in an intelligent manner.

From *Nature* 4 October 1900.

50 YEARS AGO

As a result of examination of the hands of decomposed cadavers over a long period, it has been noticed that as a rule the muscles and skin of the left hand exhibit signs of greater decomposition than the right. This may be explained by the fact that the muscles, tendons, aponeuroses, etc., are less developed in the left hand than in the right, and would, therefore, less effectively withstand the ravages of the elements. There may be other causes for the more rapid decomposition of the left hand; but it would need confirmation by the examination of the decomposed hands of left-handed persons and of the ambidextrous. Unfortunately, it is difficult to obtain reliable data for this purpose. Since the left hand appears to be more susceptible to change under post-mortem conditions than the right, it was thought that the left hand of a living person might show certain signs indicative of disease, or even the approach of disease. A study of finger prints of persons suffering from certain pathological conditions lends some support to this contention.

From *Nature* 7 October 1950.

co-workers^{3,4} found that computer simulations resulted in unexpectedly aggressive transfer of materials from resonances. First, most asteroids and meteorites fell into the Sun, requiring an even larger supply from the asteroid belt to explain the smaller fraction that reaches Earth. Second, transfer time-scales were unexpectedly short — just a few million years, incompatible with the tens of millions of years since the meteorites were first exposed to cosmic rays, known as their cosmic-ray exposure age. The amount of cosmic-ray exposure indicates the time that has lapsed since the meteorites were blasted off their parent asteroids. Mysteriously, most of the exposure must have occurred while the meteorite was in transit to the resonance.

Farinella and colleagues⁵ then turned to a long-dormant idea involving the effect of asymmetric emission of thermal radiation on the movement of small objects. Essentially, the 'afternoon' side of a rotating body is warmer than the 'morning' side, so the re-radiation of the Sun's heat produces a small net force across its path, causing the asteroid to drift from its original orbit. In addition to this diurnal effect, there is an analogous seasonal asymmetry. This effect was discovered in the mid-1960s by a graduate student, Charles Peterson, who applied it to the control of small spacecraft. Later, Peterson published a paper in *Icarus*⁶, which considered a two-step mechanism involving both resonances and asymmetric thermal forces to explain the meteorites' cosmic-ray exposure ages.

Peterson's paper was reviewed by the elderly astronomer Ernst Öpik, who had never forgotten a largely ignored explanation of such thermal forces published by a Russian engineer, I. O. Yarkovsky, in about 1900.

Peterson's work on the relevance of the Yarkovsky effect to meteorites also remained unappreciated until its recent resurrection by Farinella. Rather than being ejected directly into resonances, metre-sized fragments from asteroids a long way from the gap can be gradually moved towards resonances, while being exposed to cosmic rays, over tens of millions of years. Such slowly acting Yarkovsky forces, followed by the chaotic ejections from resonances, might together account for cosmic-ray exposure ages and deliver copious meteorites to Earth. But the idea needed quantitative proof. That is what Vokrouhlický and Farinella now provide¹. Their computer simulations of the Yarkovsky-influenced orbits of tens of millions of asteroidal fragments simultaneously track their collisions — by which the asteroids are fragmented into smaller pieces — while delivering the fragments to the two major resonances for rapid transfer to the Earth. The authors ran simulations for Hebe as well as for two asteroids far from the resonances: the inner-belt asteroid Flora, and Vesta, the putative parent body of the so-called achondritic meteorites.

The combination of Yarkovsky and collisional processes results in a surprisingly high efficiency of about 0.5% for delivery of asteroidal material to Earth, although most still falls into the Sun or ends up elsewhere. The simulated cosmic-ray exposure ages agree with meteoritic measurements. So we finally know that our collected meteorites represent asteroids from throughout the inner half of the main asteroid belt, rather than just a few, specially located bodies. Indeed, there must be many ordinary chondritic asteroids, as meteoriticists always assumed; the spectroscopic mismatch was a red herring, now

Bose-Einstein condensates

The next big thing

Simple diatomic molecules, such as oxygen (O_2), have a uniform charge distribution and an interatomic distance of a few ångströms ($1 \text{ Å} = 10^{-10} \text{ m}$). In a report in *Physical Review Letters* (85, 2458–2461; 2000), Chris Greene and colleagues predict that a vastly different diatomic molecule, Rb_2 , could exist in a quantum gas of rubidium atoms — a Bose-Einstein condensate (BEC).

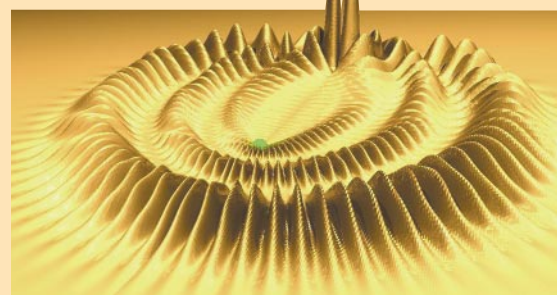
If a rubidium atom with a loosely attached electron forms in a BEC, it could bind with a normal Rb ground state atom, forming an extremely large molecule with an interatomic

distance of 500–50,000 Å. The molecule also has a unique charge distribution, shown here, reminiscent of a trilobite fossil. The Rb ion is the green sphere; the other atom is below the twin peaks.

The asymmetry of the charge distribution indicates a

large electric dipole. If this unusual molecule can be formed, its dipole will be easily manipulated by small electric fields, unlike normal molecular dipoles, which require much larger (and more expensive) fields.

Josette Chen



thought to be caused by superficial discolorations of asteroid surfaces following exposure to ions in the solar wind and bombardment by tiny meteorites.

Although some details remain to be worked out, such as the link between meteorite types and specific asteroidal parents, the theory for how meteorites get to Earth is basically complete. It is a fitting accomplishment for the final paper co-authored by Paolo Farinella, who tragically succumbed in

March to a heart problem that had weakened him since childhood.

Clark R. Chapman is at Southwest Research Institute, Suite 426, 1050 Walnut Street, Boulder, Colorado 80302, USA.

e-mail: cchapman@boulder.swri.edu

1. Vokrouhlický, D. & Farinella, P. *Nature* **407**, 606–608 (2000).
2. Wisdom, J. *Nature* **315**, 731–733 (1985).
3. Farinella, P. *et al.* *Nature* **371**, 314–317 (1994).
4. Gladman, B. J. *et al.* *Science* **277**, 197–201 (1997).
5. Farinella, P. *et al.* *Icarus* **132**, 378–387 (1998).
6. Peterson, C. *Icarus* **29**, 91–111 (1976).

tained in mice that lack follicular dendritic cells⁷ and in mice deprived of helper T cells⁸. This shows that maintenance of memory B cells may not need antigen depots or signals from helper T cells. Moreover, immunological memory also occurs in the cytotoxic T cells that destroy virus-infected cells. Here, too, memory has been shown to persist under experimental conditions in which antigen is effectively absent^{9,10}. This antigen-independent persistence of memory T cells may be a general theme of immunity and so apply to B cells too.

In their study, Maruyama *et al.*¹ used an inducible genetic switch to generate memory B cells in the absence of their specific antigen. The experiments involved genetically modifying mice to carry two different antibody genes specific for two structurally unrelated antigens. Antibody gene expression caused about 4% of the normal B cells in these mice to express specificity for the first antigen, nitrophenyl. However, although they had increased in frequency compared with an unmanipulated mouse, the nitrophenyl-specific B cells were still 'naïve', having never encountered antigen before. The antibody gene specific for the second antigen, phycoerythrin, was held in an inactive form within each nitrophenyl-

Immunology

Memory needs no reminders

Stephen Martin and Chris Goodnow

Your body can 'remember' specific diseases and, decades later, protect you from reinfection. This principle has been known for centuries, but immunologists are still struggling with the cellular mechanisms that underlie its maintenance. Is this long-term immune 'memory' due to the formation of a special class of long-lived memory cells, or to occasional 'reminders' to the immune system from residual antigens — the invading agents that induced the response in the first place. On page 636 of this issue¹ Maruyama and colleagues describe their use of an elegant genetic system to address the question. They suggest that B cells, the class of immune cell that makes antibodies, can persist as memory cells for long periods of time after immunization, even in the total absence of specific stimulation by antigen.

To explain immune memory, Burnet and Talmage² proposed that rare antigen-specific B-lymphocytes, which are precursors of antibody secreting cells, expand clonally when they encounter antigen. This clonal expansion results in an enlarged pool of antigen-specific B cells that produces a faster, larger response the next time the antigen is encountered — that is, they carry a memory of the original immunization. However, there was then a divergence of ideas about how the expanded pool of cells could persist, and this divide of opinion remains.

One line of thought proposes that intermittent re-stimulation of the memory B-cell population by specific or cross-reactive antigen is required. In the late 1960s it was discovered that intact protein antigen could exist in lymphoid tissues for months after immunization, in the form of antigen-antibody complexes on so-called follicular dendritic cells^{3,4}. These immune complexes were thought to be an antigen depot for the continual stimulation of clonally expanded B cells⁵ and for helper T cells, which aid in priming the antigen-specific B cells to produce antibodies. In support of this view,

cell-transfer experiments⁶ showed that functional B-cell memory declines rapidly with a half-life of around 3 weeks in the absence of antigen.

The contrasting opinion holds that memory B cells represent a specialized state of long-lived quiescent cells that do not require further stimulation by antigen. Supporting evidence for this comes from the finding that B-cell memory can be main-

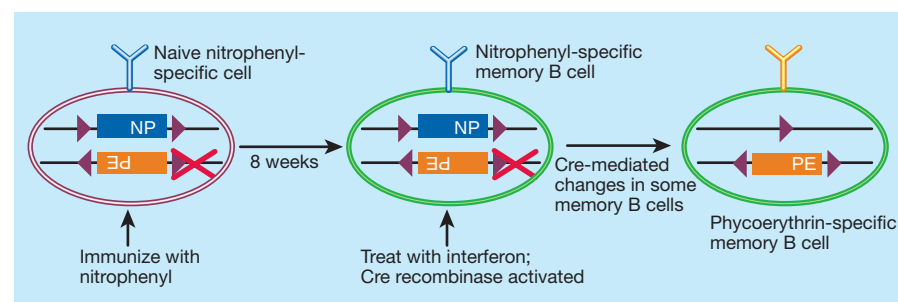


Figure 1 Does immune memory need reminders from antigen? This is the genetic switch used by Maruyama *et al.*¹ to alter the specificity of memory B cells by genetic recombination. The antigens involved were nitrophenyl (NP) and phycoerythrin (PE). After the induction of nitrophenyl-specific memory B cells, treatment with interferon produced some phycoerythrin-specific memory B cells in the absence of their specific antigen. The lifespans of the nitrophenyl- and phycoerythrin-specific memory B cells were then compared.

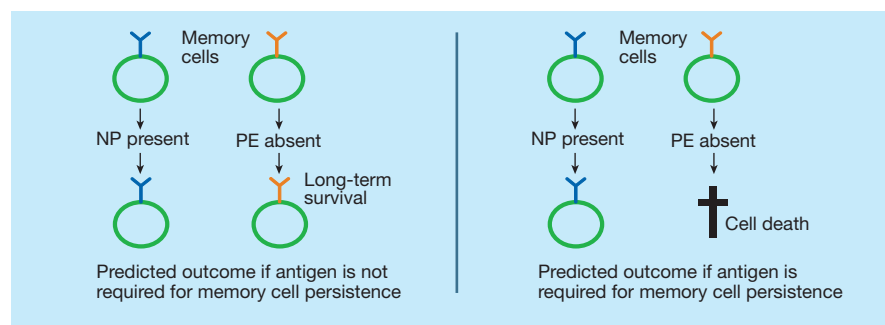


Figure 2 The predicted experimental outcomes of Maruyama and colleagues' experiments¹. The only antigen used in this experiment was nitrophenyl; phycoerythrin was completely absent. It follows that if a specific antigen is required for the maintenance of memory B cells, the phycoerythrin-specific memory B cells will die; if a specific antigen is unnecessary for persistence, these cells will survive. The authors found that the phycoerythrin-specific memory B cells persist just as well as the nitrophenyl-specific cells, and conclude that a specific antigen is not required for the maintenance of memory B cells.

specific B cell. These two antibody genes were flanked by *loxP* sites, which allowed the genes to be modified by an enzyme called Cre recombinase. The recombinase gene itself was under inducible control — Cre recombinase was made only when the mice were injected with the inducer, type-1 interferon.

Maruyama *et al.* immunized the mice with nitrophenyl to generate nitrophenyl-specific memory B cells. Memory B cells differ from naive B cells in that they are the 'antigen-experienced' progeny of the original naive B cells and are thought to be responsible for immune memory. After 8 weeks, when a pool of nitrophenyl-specific memory B cells had been established, the mice were treated with interferon. This activated the Cre recombinase, which, in a small proportion of memory B cells, caused the irreversible loss of nitrophenyl specificity and a switch to phycoerythrin specificity (Fig. 1). The phycoerythrin-specific memory B cells were predicted to persist if antigen is not required for memory-cell maintenance, or to die if antigen is essential (Fig. 2). The upshot was that these cells were maintained just as well as the nitrophenyl-specific memory B cells for 15 weeks after treatment with interferon.

Do these results rule in favour of the antigen-independent camp? The answer depends on how quickly one would expect the phycoerythrin-specific memory cells to die off in the absence of antigen, if antigen is indeed required for memory-cell persistence. The rapid decline of memory in cell-transfer experiments⁶ suggests that memory cells would die off with a half-life of around 3 weeks. Maruyama and colleagues' results clearly show that phycoerythrin-specific memory cells persist for much longer than that. But the apparent contradiction between the two sets of findings^{1,6} may have its roots in the unusual experimental design of the cell-transfer experiments — by sampling B cells from thoracic duct lymph, as opposed to sampling non-recirculating B cells from the spleen, the study may have measured the short lifespan of activated B cells rather than that of memory cells.

The more relevant question for Maruyama *et al.* is whether or not the phycoerythrin-specific memory cells die off more slowly than the phycoerythrin-specific naive B cells in the absence of antigen. Naive B cells have been shown to have a wide variety of lifespans, with an average half-life extrapolated to be at least 6–12 weeks^{11–13}, and many cells lasting for much longer. Given the potentially long lifespan of naive phycoerythrin-specific cells, a clear answer to whether antigen is needed to maintain pools of memory B cells must await experiments that monitor the decay of both naive and memory phycoerythrin-specific cells over longer periods, say 6–12 months.

The new findings¹ raise a challenging question for immunologists. If memory B cells do not require persistent immunizing antigen to survive for months or years, what stops large numbers of memory cells accumulating over a lifetime of different infections?

Stephen Martin and Chris Goodnow are at the Medical Genome Centre, John Curtin School of Medical Research, The Australian National University, Mills Road, PO Box 334, Canberra, ACT 2601, Australia.

e-mails: stephen.martin@anu.edu.au
chris.goodnow@anu.edu.au

Microbiology

Resolving a methane mystery

Edward F. DeLong

Large reservoirs of natural methane, constituting perhaps twice the amount of all known fossil-fuel stores, lie buried beneath the sea floor. Yet little of this vast methane reserve regularly escapes from oxygen-depleted (anoxic) marine sediments into the surrounding water column. Why? Microbes in marine sediments are thought to consume the methane that diffuses upwards, before it can escape into oxygenated waters. Yet no microorganism has yet been specifically identified that can make a living by oxidizing methane in the absence of oxygen. On page 623 of this issue, however, Boetius *et al.*¹ provide a pictorial view of some microbial partners involved in this process. These include members of two of the three domains of life: methane-consuming archaea and sulphate-reducing bacteria.

We need a better understanding of the methane cycle for two main reasons. First, methane is an important energy source, representing a clean and potentially economic alternative fuel. Interest in methane has recently heightened with the discovery of enormous methane stores off continental margins. These vast reservoirs, buried deep within marine sediments, exist largely in the form of a crystalline, frozen mixture of methane and water known as methane hydrate². Second, methane is a powerful greenhouse gas which, molecule for molecule, is about 25 times more potent than CO₂ in its radiative effect³. This is especially pertinent now that the world faces the prospect of global warming. But it is also relevant to Earth's climatic history: in the past, periodic outgassing from sub-seafloor hydrates may have contributed to millennial-scale oscillation of methane levels in the atmosphere⁴.

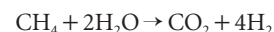
Microbes are the key to understanding the methane cycle on Earth. Most atmospheric methane appears to be microbially

1. Maruyama, M., Lam, K.-P. & Rajewsky, K. *Nature* **407**, 636–642 (2000).
2. Burnet, F. M. *The Clonal Selection Theory of Acquired Immunity* (Cambridge Univ. Press, London, 1959).
3. Nossal, G. J. V. & Ada, G. L. *Antigens, Lymphoid Cells and the Immune Response* (Academic, New York, 1971).
4. Tew, J. G. & Mandel, T. E. *Immunology* **37**, 69–76 (1979).
5. Tew, J. G., Kosko, M. H., Burton, G. F. & Szakal, A. K. *Immunol. Rev.* **117**, 185–211 (1990).
6. Gray, D. & Skarvall, H. *Nature* **336**, 70–73 (1988).
7. Karrer, U. *et al.* *J. Immunol.* **164**, 768–778 (2000).
8. Vieira, P. & Rajewsky, K. *Int. Immunol.* **2**, 487–494 (1990).
9. Murali-Krishna, K. *et al.* *Science* **286**, 1377–1381 (1999).
10. Swain, S. L., Hu, H. & Huston, G. *Science* **286**, 1381–1383 (1999).
11. Sprent, J. & Basten, A. *Cell. Immunol.* **7**, 40–59 (1973).
12. Ron, Y. & Sprent, J. *J. Exp. Med.* **161**, 1581–1586 (1985).
13. Förster, I. & Rajewsky, K. *Proc. Natl Acad. Sci. USA* **87**, 4781–4784 (1990).

generated, originating from methane-producing (methanogenic) archaea that live in anaerobic pockets around the globe. Methanogenesis is a natural process, but it can also be boosted through human agency — for example on farms in the rumen of cows, or in emanations from rice paddies. On the other side of the cycle are the methanotrophic (methane-consuming) microbes. All of the well known methanotrophs are aerobic bacteria, requiring oxygen to use methane. They are all found in oxygen-rich habitats.

However, geochemists have long suspected that methane oxidation also occurs in anoxic marine sediments. In a variety of such settings, sulphate becomes depleted downwards through the sediments, matched by an upward depletion of methane — and the minima of the two depletion curves intersect⁵ (Fig. 1, overleaf). Furthermore, isotopic evidence indicates that the CO₂ found near this boundary is derived largely from methane⁶. Radiotracer⁷ and metabolic-inhibition experiments⁸, as well as stable carbon-isotope fractionation studies⁹, have provided further detail. From these data it appears that methane oxidation in the sediments is coupled to the reduction of sulphate, an abundant oxidant in sea water.

Preliminary studies have indicated that some methanogens have limited capacity to reverse their normal metabolism, thereby oxidizing methane, rather than producing it from H₂ and CO₂¹⁰. Such 'reverse methanogenesis'



would be energetically favourable if the H₂ end product were rapidly removed, and so kept at a low steady-state concentration¹¹. In this scheme, H₂, an end product of the methane-oxidizing microbes, becomes the energy

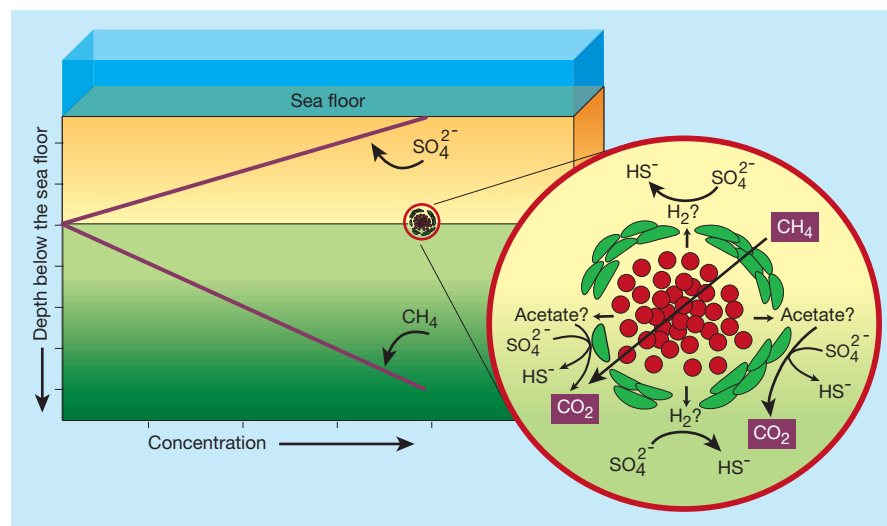
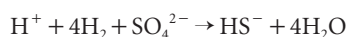


Figure 1 Methane and sulphate concentration profiles in anoxic marine sediments (after ref. 5). The sulphate-methane interface, where minima for both methane and sulphate intersect, is due presumably to anoxic methane consumption coupled to sulphate reduction. The inset shows the spatial arrangement, and potential metabolic couplings, of the microbial consortium reported by Boetius *et al.*¹. The red cells inside the aggregate are methane-consuming archaea; the green cells on the periphery are sulphate-reducing bacteria.

substrate for another microbial group, the hydrogen-oxidizing sulphate-reducers:



The sum of this cooperative process between methane-oxidizers and sulphate-reducers is



This type of intermicrobial metabolic interaction, termed 'syntrophy', drives several metabolic transformations, and may even have been central to the evolution of cellular organelles¹². Exactly which metabolites are being exchanged between the partners remains uncertain. Evidence from stable-isotope biomarkers indicates that methane-derived carbon is efficiently transferred from methane-oxidizers to sulphate-reducing bacteria¹³ (perhaps via acetate; Fig. 1, inset). It seems that hydrogen or acetate, or both, are the probable metabolites that shuttle energy, and possibly carbon, from the methane-oxidizers to their sulphate-reducing partners.

So, which microbes are responsible for methane oxidation in the absence of oxygen? There is good evidence that archaea closely related to methanogens are involved. Recent analyses of microbial communities associated with methane seeps showed that archaeal lipids there were derived directly from methane; and phylogenetic work on the same samples identified just a few archaeal groups as the likely anaerobic consumers of methane¹⁴. This study also provided the data necessary to design the fluorescently labelled 'phylogenetic stains'¹⁵ used by Boetius *et al.*¹.

Using these stains on samples of methane-rich marine sediment, Boetius *et al.* could identify the putative methane-oxidizing archaea microscopically. They provide visual evidence that one previously identified sediment-associated archaeal group¹⁴, affiliated with the order Methanosarcinales, is a key methane-oxidizer. Moreover, their work shows that the organisms are spatially organized in tight clusters, closely surrounded by their sulphate-reducing bacterial partners (Fig. 1, inset). This is the strongest evidence yet that methane oxidation in anoxic marine sediments is carried out by a consortium of archaeal methane-consumers and bacterial sulphate-reducers. And the spatial co-location of archaea and bacteria suggests just how the syntrophic interaction might be coupled.

Precision measurements

Optical clocks coming of age

Patrick Gill

The caesium atomic clock was introduced some 45 years ago, and adopted as the international definition of time in 1967. The basis of the standard is the absorption of microwave radiation at 9.2 gigahertz (GHz) by caesium atoms. Its measurement precision is determined primarily by the narrowness or spectral spread of the absorption in relation to its frequency. Physicists have since speculated that optical absorptions (in the near infrared, visible or ultraviolet) would make better frequency standards and clocks^{1,2}, because of their much higher fre-

Questions remain, of course. What specific metabolites are shuttled between the archaea and their sulphate-reducing partners? Can the methane-oxidizing archaea still generate methane, or are they evolutionarily committed to unidirectional methane consumption? Are the microbes identified by Boetius *et al.* the sole players, or do other microbial groups catalyse the same process? Finally, in a tight consortium such as this, has lateral gene transfer accelerated the evolution of metabolic pathways and processes? (Methanogenic archaea and aerobic, methylotrophic bacteria have swapped genes in the past¹⁶.) With their micrographs, Boetius *et al.*¹ have added a new piece to the puzzle. But there are still plenty of missing pieces that geologists, geochemists, microbiologists and ecologists need to find.

Edward F. DeLong is at the Monterey Bay Aquarium Research Institute, 7700 Sandholdt Road, Moss Landing, California 95039, USA.
e-mail: delong@mbari.org

1. Boetius, A. *et al.* *Nature* **407**, 623–626 (2000).
2. Kvenvolden, K. A. *Chem. Geol.* **71**, 41–51 (1988).
3. Kiehl, J. T. & Dickinson, R. E. *J. Geophys. Res.* **92**, 2991–2998 (1987).
4. Kennett, J. P., Cannariato, K. G., Hendy, I. L. & Behl, R. J. *Science* **288**, 128–133 (2000).
5. Borowski, W. S., Paull, C. K. & Ussler, W. *Mar. Geol.* **159**, 131–154 (1999).
6. Borowski, W. S., Paull, C. K. & Ussler, W. *Mar. Chem.* **57**, 299–311 (1997).
7. Iversen, N. & Jørgensen, B. B. *Limnol. Oceanogr.* **30**, 944–955 (1985).
8. Alperin, M. J. & Reeburgh, W. S. *Appl. Environ. Microbiol.* **50**, 940–945 (1985).
9. Alperin, M. J., Reeburgh, W. S. & Whiticar, M. J. *Glob. Biogeochem. Cycles* **2**, 279–288 (1988).
10. Zehnder, A. J. B. & Brock, T. D. *J. Bacteriol.* **137**, 420–432 (1979).
11. Hoehler, T. M., Alperin, M. J., Albert, D. B. & Martens, C. S. *Glob. Biogeochem. Cycles* **8**, 451–463 (1994).
12. Martin, W. & Müller, M. *Nature* **392**, 37–41 (1998).
13. Hinrichs, K., Summons, R. E., Orphan, V., Sylva, S. P. & Hayes, J. M. *Org. Geochem.* (in the press).
14. Hinrichs, K., Hayes, J. M., Sylva, S. P., Brewer, P. & DeLong, E. F. *Nature* **398**, 802–805 (1999).
15. DeLong, E. F., Wickham, G. S. & Pace, N. R. *Science* **243**, 1360–1363 (1989).
16. Chistoserdova, L., Vorholt, J. A., Thauer, R. K. & Lidstrom, M. E. *Science* **281**, 99–102 (1998).

quency (approaching 1 million GHz). Essentially, if your clock ticks more quickly in a given time, your precision for time measurement improves with the tick frequency. Until recently, however, observing very narrow optical absorptions with enough spectral resolution has not been practical.

But in a paper in *Physical Review Letters*, Rafac *et al.*³ at the National Institute of Standards and Technology (NIST) report the observation of a 6.7-Hz experimental linewidth for the absorption of ultraviolet light by a single mercury (Hg⁺) ion. This is

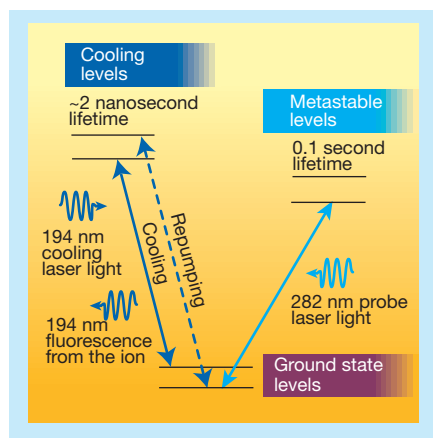


Figure 1 Energy levels of the Hg^+ ion. The ion is first cooled to 1.7 mK by the repeated absorption and re-emission of 194-nm light, when the cooling laser frequency is set just below the centre frequency of the absorption. A small amount of 'repumping' by laser light at a similar wavelength is necessary to ensure that the ion ends up at the right ground-state level. After cooling, the 282-nm light from the probe laser is used to drive the ion between the ground and a metastable state, causing the cooling fluorescence to turn off while the ion is 'shelved'. The variation in the number of quantum jumps with probe frequency is used to observe the linewidth of the narrow absorption in the ion's spectrum. The linewidth observed by Rafac *et al.*³ is the narrowest optical absorption observed to date.

close to the intrinsic quantum limit of the Hg^+ absorption. With it comes the possibility for optical clocks surpassing even the best caesium clocks.

First, the NIST group had to trap and cool their mercury ion. They ionized a mercury atom by electron bombardment under ultra-high-vacuum conditions, and immediately trapped it within an electrodynamic potential well. Once trapped, the ion can quickly be laser-cooled to temperatures around 1 mK above absolute zero by the strong absorption and re-emission of ultraviolet light at a wavelength of 194 nm. This laser-cooling can be likened to the repeated scattering of ping-pong balls (the photons) off a moving billiard ball (the ion) to reduce its momentum. The ion scatters millions of these ultraviolet photons per second, and a small fraction of this fluorescence is detected.

Once cold, the ion is virtually at rest in the trap, so there is little thermal motion to broaden the spectral linewidth of the ion's absorption. This is crucial as the particular absorptions that make good clocks have such narrow linewidths that their true width remains hidden until this broadening due to thermal motion is removed. In the single-ion system, the clock absorptions correspond to transitions from the ground-state electronic energy of the ion to a metastable state, and are barely allowed under quantum mechanical rules. In Hg^+ , the lifetime of the

metastable state is ~ 0.1 second (Fig 1), and the corresponding clock transition at 282 nm has a quantum-limited absorption spectral width of ~ 1.7 Hz. When you consider the frequency of the 282-nm absorption (that is, 1.06×10^{15} Hz), you get an idea of how narrow this absorption is.

Because this clock transition is only weakly allowed, one might expect the observed absorption signal to be very small. Here, though, quantum mechanics lends a hand. A probe laser at 282 nm is used to drive the ion to its metastable state, where it languishes for its natural lifetime, before decaying back down to the ground state (Fig. 1). While this is going on, cooling photons at 194 nm are not being scattered by the ion. The disappearance of cooling fluorescence, when the ion is 'shelved' in the metastable level, followed by its reappearance as the ion decays back down, is called a 'quantum jump', and allows the NIST group to efficiently detect the weak absorption by the single ion. By recording the number of quantum jumps per step, as the probe laser frequency is stepped across the absorption, Rafac *et al.*³ build up the spectral profile of the clock transition. For best experimental resolution, the laser linewidth and the step width need to be smaller than the width of the clock absorption. With the profile established, the probe laser light is then stepped repeatedly back and forth between the half-intensity points on either side of the profile. Any mismatch in jump rate at these points is fed back to the laser in order to hold it centred on the clock absorption frequency. This stabilized optical frequency then forms the basis of the optical clock.

The NIST group first demonstrated this nearly 10 years ago⁴, by obtaining an 80-Hz absorption linewidth in a cold ion. This remained a benchmark for other researchers to strive for, until their recent 7-Hz result. So what has led to this improvement? First, they did their latest experiment in a liquid-helium cryostat, to avoid loss of the ion by chemical reaction with background mercury atoms. Second, the NIST group have developed a laser with an extremely narrow linewidth of ~ 0.1 Hz for probing the clock transition⁵. This is a factor of 200 better than their lasers of a decade ago, and was achieved by pushing laser noise-suppression techniques to the limit.

Finally, Rafac *et al.* optimized the individual probe times at each frequency step to ~ 0.1 s, close to the natural lifetime of the clock transition. If this probe time is too short, then interrogation time broadening masks the natural width. Conversely, it is pointless for the interrogation time to be much greater than the natural decay time.

With this technique, Rafac *et al.* observe a 6.7-Hz linewidth for the clock transition at 282 nm, corresponding to a fractional frequency resolution of 6.3 parts in 10^{15} , which

Daedalus

David Jones

David Jones, author of the Daedalus column, is indisposed.

is the narrowest absorption observed within the optical region to date. There remains much to learn, especially about the extent to which environmental perturbations, such as magnetic and electric fields, disturb the ion's clock frequency. But even the preliminary NIST data indicate a frequency uncertainty that is not much worse than the best atomic clocks, which are now at 1 part in 10^{15} . Improvements in the control of magnetic field and trap geometry should allow the NIST system to reach an uncertainty of 1 part in 10^{18} .

This work is a clear precursor to the arrival of the single-ion optical clock. New developments in femtosecond laser technology^{6,7} provide a much simplified and accurate way to 'count' high-frequency optical cycles in terms of cycles of the caesium microwave standard. Maybe we should now start to think of reversing this process and redefining the second in terms of an optical clock. Is mercury the best choice for such an ion clock? There are other similar single-ion systems, such as indium⁸, ytterbium^{9,10}, strontium^{11,12}, barium¹³ and calcium¹⁴, some of which have clock transitions with smaller natural widths, so this is still an open question.

But we should not forget that caesium clocks operate successfully at the 1 part in 10^{15} level, and rubidium microwave clocks have the potential to improve on this^{15,16}. In general, better clocks will have an important impact on global and satellite positioning and timing capabilities, as well as future communication networks. They also help to increase our knowledge of atomic structure, as well as enabling more accurate measurement of fundamental physical constants, such as whether they vary with time, and more precise tests of quantum physics. ■

Patrick Gill is at the National Physical Laboratory, Teddington, Middlesex TW11 0LW, UK.

e-mail: patrick.gill@npl.co.uk

1. Dehmelt, H. G. *IEEE Trans. Instrum. Meas.* **31**, 83–87 (1982).
2. Wineland, D. J. *Science* **226**, 395–400 (1984).
3. Rafac, R. J. *et al. Phys. Rev. Lett.* **85**, 2462–2465 (2000).
4. Bergquist, J. C. *et al. in Proc. Int. School of Physics Enrico Fermi CXX 359* (North Holland, Amsterdam, 1994).
5. Young, B. C. *et al. Phys. Rev. Lett.* **82**, 3799–3802 (1999).
6. Reichert, J. *et al. Phys. Rev. Lett.* **84**, 3232–3235 (2000).
7. Diddams, S. A. *et al. Phys. Rev. Lett.* **84**, 5102–5105 (2000).
8. Zanthier, J. V. *et al. Opt. Commun.* **166**, 57–63 (1999).
9. Tamm, C. *et al. Phys. Rev. A* **61**, 053405 (2000).
10. Roberts, M. *et al. Phys. Rev. A* **62**, 020501(R) (2000).
11. Bernard, J. E. *et al. Phys. Rev. Lett.* **82**, 3228–3231 (1999).
12. Barwood, G. P. *et al. Phys. Rev. A* **59**, R3178–R3181 (1999).
13. Yu, N. *et al. Phys. Rev. A* **50**, 2738–2741 (1994).
14. Urabe, S. *et al. Appl. Phys. B* **67**, 223–227 (1998).
15. Fertig, C. & Gibble, K. *Phys. Rev. Lett.* **85**, 1622–1625 (2000).
16. Sortais, Y. *et al. Phys. Rev. Lett.* (in the press).

Colour-enhancing protein in blue petals

Spectacular morning glory blooms rely on a behind-the-scenes proton exchanger.

The coloration of blue flowers depends on the production of the appropriate anthocyanin pigments, the presence of metal ions and co-pigments, and the vacuolar pH¹. An increase in vacuolar pH enhances blue coloration^{1,2}, but little is known about the proteins responsible for raising the vacuolar pH. Here we show that a mutant purple-flowering Japanese morning glory (*Ipomoea nil*) plant that carries a recessive mutation in the gene encoding a vacuolar Na⁺/H⁺-exchanger is unable to increase its vacuolar pH to create the normal bright blue petals.

In the morning glory *Ipomoea tricolor*, reddish-purple buds unfurl into blue flowers. As the bloom opens, the vacuolar pH in the flower epidermis increases from 6.6 to 7.7 (ref. 2). The colour is very similar to that of *I. nil*, and the flowers of both plants contain the same anthocyanin as their principal pigment^{2,3}.

Ipomoea nil plants bearing the purple-mutable (*pr-m*; Fig. 1a) allele in the Purple (*Pr*) locus have purple flowers with blue sectors⁴ (Fig. 1a) owing to recurrent somatic mutation from the recessive *pr-m* to the blue Purple-revertant (*Pr-r*; Fig. 1b) allele. We have produced germinal revertants (heterogenotes, *Pr-r/pr-m*) that have blue flowers (Fig. 1b) and, after self-pollination, we chose pairs of siblings carrying either the *pr-m* or *Pr-r* allele homozygously. We found no difference in the pigment composition of the *pr-m* and *Pr-r* lines, but the pH of the sap from the *Pr-r* flower epidermis was higher than that in *pr-m* by about 0.7, indicating that the *pr* mutant fails to increase the vacuolar pH.

As *Tpn1*-related transposons are a common spontaneous mutagen in *I. nil*⁵, we used a transposon-display procedure⁶ to identify the *Pr* gene. Sequencing of *Pr* complementary DNA (accession no. AB033989) revealed that it contains an open reading frame that is highly homologous to the *Arabidopsis* and rice vacuolar Na⁺/H⁺-exchangers^{7,8} and that it complements the yeast *nhx1* mutation⁷ (our unpublished results), confirming that the *Pr* gene encodes a vacuolar Na⁺/H⁺-exchanger. We named this transporter *InNhx1* (for *I. nil* Nhx1).

The *Pr* gene itself (accession no. AB033990) comprises 15 exons and a *Tpn1*-related element, *Tpn4*, which is integrated into its first exon in the *pr-m* lines (Fig. 1c). This integration of *Tpn4* generates a 3-base-pair target duplication, and the germinal revertants studied here contain footprints generated by *Tpn4* excision (Fig. 1d). Moreover, blue sectors of the variegated flowers

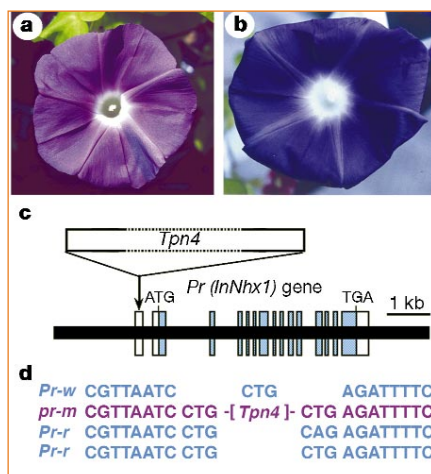


Figure 1 *Ipomoea nil* flower phenotypes and the structure of the *Pr* gene. **a**, Mutable line (*pr-m*). **b**, Germinal revertant (*Pr-r*). **c**, The *Pr* gene. White and blue boxes represent the *Pr* exons and open reading frame, respectively. **d**, Sequences at the *Tpn4* insertion site in the *Pr-r* (wild type), *pr-m* and *Pr-r* lines.

contain footprints that result from somatic excision of *Tpn4*, whereas we found no evidence of such excision in the purple regions of these flowers in six different *pr-m* lines.

Materials science

Ti₃SiC₂ has negligible thermopower

A part from superconducting materials below their critical transition temperature, all other materials subjected to a temperature gradient will develop an electromotive force or (absolute) thermopower which is itself a function of temperature. We have discovered that the thermopower of Ti₃SiC₂ is essentially zero over an extended temperature range (from 300 to 850 K). This material should allow the thermopower of other substances to be determined directly at high temperatures, a measurement that has so far been impossible.

Characterization of single-phase, bulk dense samples of Ti₃SiC₂ has revealed that this ternary carbide possesses a unique combination of properties^{1–5}. It has high elastic moduli and easy machinability¹, good thermal and electrical conductivity^{2,3}, together with excellent oxidation properties⁴ and fatigue resistance⁵, as well as being tough and tolerant to damage⁵. We have now investigated another property of this remarkable material, namely its absolute thermopower, Θ , over the temperature range 300 to 850 K.

To our knowledge, the *Pr* gene encoding *InNhx1* is the first gene to be identified that increases the vacuolar pH in petals in order to confer blue coloration.

Sachiko Fukada-Tanaka*, Yoshishige Inagaki*, Toshio Yamaguchi†, Norio Saito‡, Shigeru Iida*†

*National Institute for Basic Biology, Okazaki 444-8585, Japan
e-mail: shigiida@nibb.ac.jp

†Department of Molecular Biomechanics, Graduate University for Advanced Studies, Okazaki 444-8585, Japan

‡Chemical Laboratory, Meiji-gakuin University, Yokohama 244-8539, Japan

- Mol, J., Grotewold, E. & Koes, R. *Trends Plant Sci.* **3**, 212–217 (1998).
- Yoshida, K., Kondo, T., Okazaki, Y. & Katou, K. *Nature* **373**, 291 (1995).
- Lu, T. S., Saito, N., Yokoi, M., Shigihara, A. & Honda, T. *Phytochemistry* **31**, 659–663 (1992).
- Imai, Y. *J. Coll. Agric. Imp. Univ. Tokyo* **12**, 479–523 (1934).
- Iida, S., Hoshino, A., Johzuka-Hisatomi, Y., Habu, Y. & Inagaki, Y. *Ann. NY Acad. Sci.* **870**, 265–274 (1999).
- Maes, T., De Keuleleire, P. & Gerats, T. *Trends Plant Sci.* **4**, 90–96 (1999).
- Gaxiola, R. A. *et al. Proc. Natl Acad. Sci. USA* **96**, 1480–1485 (1999).
- Fukuda, A., Nakamura, A. & Tanaka, Y. *Biochim. Biophys. Acta* **1446**, 149–155 (1999).

We measured the absolute thermopower and electrical resistivity, ρ , simultaneously on the same specimen of single-phase Ti₃SiC₂ in the 300–850 K temperature range, as described^{6,7}. The 4-probe d.c. resistivity of

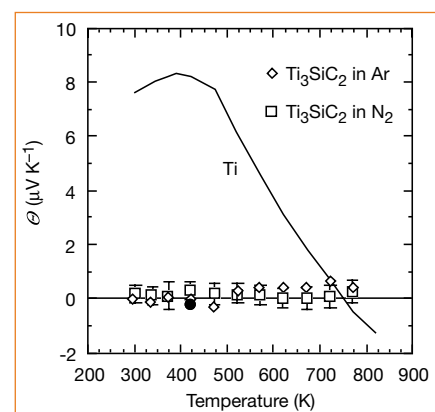


Figure 1 Temperature dependence of thermopower, Θ , of Ti₃SiC₂. Two different techniques were used to confirm the reproducibility of the data⁶: a heat pulse and steady state (solid circle at 423 K) in two different atmospheres (nitrogen or argon). Each datum point was measured at least ten times and the error bar represents one standard deviation, which was always less than 0.3 $\mu\text{V K}^{-1}$. The thermopower as measured was subsequently corrected against the absolute thermopower of the platinum leads¹² to obtain the absolute thermopower of Ti₃SiC₂. The curve representing Θ of polycrystalline Ti is from ref. 8.

Ti₃SiC₂ is metallic-like and can best be represented against temperature T (in degrees Kelvin) as $\rho = (0.224 \pm 0.002)[1 + (0.00371 \pm 0.0003)(T - 300)] \mu\Omega \text{ m}$, which is lower than that of titanium metal and in good agreement with previous results⁵.

The temperature dependence of Θ is shown in Fig. 1. The data are in good agreement with each other, irrespective of measurement technique or atmosphere, supporting their reproducibility and reliability. The absolute thermopower may best be estimated as $\Theta = 0.18 \pm 0.22 \mu\text{V K}^{-1}$.

This shows that Θ is practically zero within $\pm 0.22 \mu\text{V K}^{-1}$ in the temperature range examined. Based on this result, we would expect Θ to remain zero even outside the 300–850 K range. Included for comparison in Fig. 1 is Θ of titanium metal⁸, which is like most other solids in that its thermopower is a function of temperature.

The reason for this vanishingly small thermopower is not entirely clear. There is little doubt, however, that like Ti and some other transition metals^{9,10}, both electrons and holes are involved in the transport properties. It is only by having oppositely charged particles that Θ can vanish¹¹. Our results imply that the thermoelectric contributions by electrons and holes delicately cancel each other out over a wide temperature range. Hall coefficient measurements indicate that it also fluctuates between positive and negative values⁷; that is, it is also approximately zero. This, together with the present results, suggests that the concentration, mobility and heat of transport values of the electrons are identical to those of the holes over the temperature range investigated.

So far, all thermopower measurements — including ours — have had to be corrected for by adding the absolute thermopower of the lead wires. Such corrections can, if not carefully made, introduce errors. The benefit of using a material with a thermopower of zero over an extended temperature range to measure the thermopower of other substances is obvious. Ti₃SiC₂ might also find an application as leads for high-temperature thermocells.

H.-I. Yoo*, M. W. Barsoum†, T. El-Raghy†

*School of Materials Science and Engineering, Seoul National University, Seoul 151-742, Korea

†Department of Materials Engineering, Drexel University, Philadelphia, Pennsylvania 19104, USA

1. Barsoum, M. W. & El-Raghy, T. J. *Am. Cer. Soc.* **79**, 1953–1956 (1996).

2. Barsoum, M. W. & El-Raghy, T. J. *Mater. Synth. Process.* **5**, 197–216 (1997).

3. Barsoum, M. W. et al. *J. Phys. Chem. Solids* **60**, 429–439 (1999).

4. Barsoum, M. W., El-Raghy, T. & Ogbuji, L. *J. Electrochem. Soc.* **144**, 2508–2516 (1997).

5. Gilbert, C. J. et al. *Scripta Mater.* **42**, 761–767 (2000).

6. Yoo, H.-I. & Hwang, J.-H. *J. Phys. Chem. Solids* **53**, 973–981 (1992).

7. Barsoum, M. W. et al. *Phys. Rev. B* (in the press).

8. Worner, H. W. *Aust. J. Sci. Instrum.* **4**, 62–72 (1951).

9. Berlincourt, T. G. *Phys. Rev.* **114**, 969–981 (1959).

10. Mott, N. F. & Jones, H. *The Theory of the Properties of Metals and Alloys* (Dover, New York, 1958).

11. Yoo, H.-I. & Kim, C.-S. *Solid State Ionics* **53–56**, 583–591 (1992).

12. Cusack, N. & Kendall, P. *Proc. Phys. Soc.* **72**, 898–901 (1958).

Biomechanics

Asymmetric toes aid underwater swimming

The unique morphology of the toes of the great crested grebe (*Podiceps cristatus*), which are asymmetrically lobed with a narrower skin flap on the lateral side of the toe, enables these birds to swim very efficiently. Here we study video recordings of a diving grebe and stroboscopic pictures of its moving feet and conclude that the bird uses a hydrodynamically lift-based foot (power) stroke to propel itself underwater, with the separated toes functioning as multiple slots to increase the lift-to-drag ratio. The asymmetric lobes are an adaptation for self-stabilization of the toes during the power stroke, and the toes themselves act as separate hydrofoils, each producing lift and each being twistable individually under hydrodynamic load.

Three-dimensional kinematics reveal that during the power stroke, the grebe's feet move through an arc upwards and medially, but not backwards relative to the still water, while the spread toes are twisted and the lateral, fourth, toe is leading the movement. A large lift force is then produced that is directed forwards — providing a large thrust component — and a smaller drag force is directed outwards and downwards. The propulsion is thus based primarily on a hydrodynamically lift-producing leg and foot stroke (our unpublished results), in contrast to the drag-based locomotion assumed previously^{1,2}.

Lift-based paddling is quicker and more energy-efficient than the drag-based propulsion used by diving birds such as

anseriforms. These swimming movements can also be compared to asymmetric hovering in flying animals³, as the feet may instantaneously accelerate a flow through the circular sweep made by the feet.

The function of the grebe's toes can be compared to that of flight feathers in several respects³. The separated toes most probably function as multiple slots, thereby increasing the lift-to-drag ratio during swimming and probably also the maximum lift⁴. This is dependent on the bird using a lift-based power stroke. Furthermore — as our kinematic analyses suggest (Fig. 1a) — the toes must be spread and staggered in height in relation to the oncoming water during the power stroke.

The asymmetric lobing of the grebe's toes seems to be an adaptation to self-stabilize the toes during the power stroke, like the primary-feather asymmetry of flying birds⁵. The asymmetry of the grebe's toes at one quarter of the distance from the tips is close to what is found in handwing feathers. In both, the axis of rotation (at the toe phalanges and feather shaft) lies at about 1/5–1/3 of the chord length from the leading edge (Fig. 1b), with the greatest asymmetry on the toe leading the movement during the power stroke (toe IV), as would be expected from the pressure distribution of multiple slots⁶.

The toes may act as separate hydrofoils, each producing lift and each twistable individually in the nose-down direction under hydrodynamic load. Separated feathers, acting as separate airfoils, can be twisted more than the entire, unseparated wing, with higher local angles of attack and lift coefficients^{3,7}.

In the grebe, the rotation of the foot produces a speed gradient along the toes, with increasing velocity from base to tip. Increasing asymmetry towards the tips of the toes may allow the grebe to twist the toes passively and keep favourable angles of attack along the entire length of the toes during the power stroke.



Figure 1 The great crested grebe's separated toes produce lift. **a**, Stroboscopic pictures of the grebe's right foot as it is moved mechanically in a still-water tank, simulating the power stroke. The foot is moved from right (first part of the power stroke) to left (last part of power stroke) in the figure, which shows five different phases. Note the staggering and twisting of the toes in the picture on the left. **b**, Dorsal view of the same foot as shown in **a**. The asymmetry of the toes is most pronounced on the fourth toe, which is leading the movement during the power stroke, and least on the second. The second and third toes in particular show increasing asymmetry towards their tips. Photographs by L.C.J.

The grebe family is an ancient group, with no close relatives among living birds. As far as we know, all grebes have asymmetrically lobed toes and the mode of swimming described here is probably optimally suited to this morphology.

L. Christoffer Johansson,

Ulla M. Lindhe Norberg

Department of Zoology, Zoomorphology, Göteborg University, 413 90 Göteborg, Sweden

- Webb, P. & Blake, R. W. in *Functional Vertebrate Morphology* (eds Hildebrand, M., Bramble, D. M., Liem, K. F. & Wake, D. B.) 110–128 (Harvard Univ. Press, Cambridge, MA, 1985).
- Braun, J. & Reif, W.-E. *N. Jb. Geol. Paläont. Abh.* **169**, 307–332 (1985).
- Norberg, U. M. *Vertebrate Flight* (Springer, New York, 1990).
- Hofton, A. *New Sci.* 146–147 (20 April 1978).
- Norberg, R. Å. *Biol. Rev.* **48**, 561–596 (1973).
- Smith, A. M. O. *J. Aircraft* **12**, 501–530 (1975).
- Norberg, R. Å. in *The Beginnings of Birds* (eds Hecht, M. K., Ostrom, J. H., Viohl, G. & Wellnhofer, P.) (Proc. Int. Archaeopteryx Conf. Eichstätt, 1984, 1985).

Germine DNA

Wheat mutation rate after Chernobyl

The accident at the Chernobyl nuclear power plant in 1986 has generated concern over the genetic consequences of chronic exposure to radiation. Here we describe a new approach to monitoring germline mutation in plants and find evidence for a remarkably strong induction of germline mutation in wheat upon chronic exposure to ionizing radiation produced by the Chernobyl accident.

We compared wheat plants descended from two genetically identical populations, derived from the same homogeneous parental line. One population was grown for one generation (10 months) in a heavily contaminated plot (900 Ci km⁻²) near the Chernobyl nuclear power plant¹, the other was sown in a clean (<1 Ci km⁻²) control area 30 km away in soil with comparable agrochemical characteristics.

Using the polymerase chain reaction, we profiled offspring plants for 13 single-copy monomorphic microsatellite loci². Evidence for alterations (variants) was obtained for all 13 loci, including gains and losses of repeats, as well as complete loss of microsatellite bands (nulls) (Table 1). Offspring derived from exposed plants showed no increase in the frequency of homozygous variants and a threefold increase in the frequency of heterozygous structural variants, attributed to all loci within this group of plants.

Differences between the two initially identical populations, presumably arising over one generation, may result from seed contamination, migration or mutation. Seed contamination is unlikely, however, because neither plot had previously con-

Table 1 Frequency of rare microsatellite variants in the offspring of wheat plants

Type of variant	Frequency*		Ratio to control	P†	Wilcoxon test	
	Control	Exposed			z	P‡
Nulls (homozygotes)	0.0056	0.0066	1.17	0.8334	0.42	0.6754
Losses (heterozygotes)	0.0010	0.0066	6.45	0.0054	2.57	0.0100
Losses (homozygotes)	0.0015	0.0029	1.88	0.5461	0.74	0.4606
Gains (heterozygotes)	0.0031	0.0087	2.82	0.0270	1.84	0.0654
Gains (homozygotes)	0.0015	0.0021	1.34	0.9722	0.41	0.6831
Heterozygotes losses + gains	0.0041	0.0153	3.73	0.0003	3.19	0.0014
Homozygotes losses + gains	0.0031	0.0050	1.61	0.4686	0.53	0.5930

We screened 186 and 150 wheat plants grown from seeds collected from exposed and control plants, respectively (further details are available from the authors). Assuming selective neutrality, the equilibrium frequency of heterozygous variants, H_e , was approximated as $H_e = 4N_e u_e / (1 + 4N_e u_e)$, where N_e and u_e are effective population size and spontaneous mutation rate, respectively³. If $N_e = 1$ for self-fertilized plants, then $u_e \approx H_e/4$. Homozygosity in the exposed group, f_e , was approximated as $f_e = [(1/2N_e) + (1 - (1/2N_e))] \times (1 - u_e)^2$, where f_{e-1} and u_e are the homozygosity of the previous generation and mutation rate in the exposed population, respectively³. If $f_{e-1} = 1 - H_e$, then $u_e \approx (H_e - H_e/2)/2$.

*Frequency was estimated as the number of variants per microsatellite locus.

†Probability of difference from the control group (Fisher's exact test, two-tailed; statistically significant values are in bold).

‡Probability of Wilcoxon test (statistically significant values are in bold).

tained wheat. Seed contamination would also have affected the frequency of all three types of variant, including nulls and other homozygous variants. Cross-pollination can also be excluded as wheat is an obligate self-pollinator, preventing the migration of pollen between neighbouring populations. We conclude, therefore, that the increased diversity in the heterozygous variants in the offspring is probably due to a radiation-induced increase in microsatellite mutation in the exposed plants.

Assuming that the control population is in equilibrium (Table 1), we estimate that the spontaneous mutation rate is 1.03×10^{-3} per locus (95% confidence interval, 0.44×10^{-3} – 2.03×10^{-3}), whereas mutation rate in the exposed group was 6.63×10^{-3} (95% confidence interval, 4.28×10^{-3} – 9.70×10^{-3}). Thus we attribute the more than threefold increase in heterozygosity in the exposed group to a more than sixfold increase in the mutation rate over the single generation of exposure to ionizing radiation.

We estimate that the wheat plants have been exposed to relatively low doses of chronic irradiation of about 0.3 Gy, with external and internal components of 0.2 and 0.1 Gy, respectively³. Theoretically, this low-level exposure should not cause such a large increase in the mutation rate, suggesting that chronic exposure to ionizing radiation has effects that are as yet unknown. Other studies have shown that chronic internal exposure is far more efficient at inducing somatic recombination than acute external exposure^{1,4}.

The estimated increase in mutation rate in the offspring of exposed plants is too high to be due to direct targeting of microsatellite loci by ionizing radiation. The wheat genome contains 16×10^9 base pairs⁵, and the mean size of the 13

microsatellite loci included in this study is about 100 base pairs. Attributing a sixfold increase in the mutation rate to direct radiation-induced DNA damage at the microsatellite loci would mean that the expected damage to the whole wheat genome was (increase of mutation rate) $\times$ (genome size)/(size of locus) $\approx 80,000$ damaging events — much higher than any experimentally derived measurements of the initial yield of radiation-induced damage to a eukaryotic cell⁶. The increase in microsatellite mutation rate in plants may therefore be better explained by a non-targeted effect of ionizing radiation elsewhere in the cell, as described for mammalian minisatellite loci^{7,8}.

Our findings raise the important issue of the genetic hazard of chronic radiation exposure to the germ line, showing that the apparent rate of induced microsatellite germline mutation is much higher than existing estimates of absorbed doses of exposure would predict. Further study is needed to analyse the genetic effects of chronic radiation exposure.

Olga Kovalchuk*, **Yuri E. Dubrova†**, **Andrey Arkhipov‡**, **Barbara Hohn***, **Igor Kovalchuk***

*Friedrich Miescher Institute, Novartis Research Foundation, PO Box 2543, CH-4002 Basel, Switzerland

e-mail: Olga.Kovalchuk@fmi.ch

†Department of Genetics, University of Leicester, Leicester LE1 7RH, UK

‡Chernobyl Scientific and Technical Center of International Research, Shkolnaya Str. 6, 255620 Chernobyl, Ukraine

- Kovalchuk, I., Kovalchuk, O., Arkhipov, A. & Hohn, B. *Nature Biotechnol.* **16**, 1054–1057 (1998).
- Roder, M. S. et al. *Genetics* **149**, 2007–2023 (1998).
- Moiseev, A. & Ivanov, V. *Directory for Dosimetry and Radiation Hygiene* (Atomizdat, Moscow, 1991).
- Kovalchuk, O. et al. *Mutat. Res.* **449**, 47–56 (2000).
- Shields, R. *Nature* **365**, 297–298 (1993).

6. Frankenberg-Schwager, M. *Radiat. Environ. Biophys.* **29**, 273–292 (1990).
7. Sadamoto, S. *et al. Int. J. Radiat. Biol.* **65**, 549–557 (1994).
8. Dubrova, Y. E., Plumb, M., Brown, J. & Jeffreys, A. J. *Int. J. Radiat. Biol.* **74**, 689–696 (1998).
9. Crow, J. F. & Kimura, M. *An Introduction to Population Genetics Theory* (Harper, New York, 1970).

Metabolic scaling

Energy constraints on carnivore diet

The energy expenditure of mammals reflects their habits and environments¹, subject to limitations associated with body size. Carbone *et al.*² combined scaling relationships to argue that large species of the mammalian order Carnivora (weighing more than 21.5 kg) do not specialize on invertebrate prey. However, many tropical mammals that feed exclusively on ants and termites are much heavier than this, often weighing up to 60–70 kg; they survive by progressively reducing their metabolic rate to below that expected from their body size. I believe that this response indicates that it is not body size that limits the determination of diet, but rather the maximal rate of energy expenditure.

The size limit for a predator exclusively dependent on invertebrate prey is not absolute. For example, the sloth bear (*Ursus ursinus*), a carnivore that can weigh as much as 145 kg and feeds extensively (but not exclusively³) on termites, was considered by Carbone *et al.* to be an outlier — but outliers should not be ignored as they may tell us that our theories are incomplete. Their analysis² fails to recognize that all scaling relationships contain biologically relevant variation, and inherent in this residual scatter are adjustments that permit a large mass in carnivores and other terrestrial mammals that consume invertebrate prey.

Large mammals (over 20 kg) that specialize in eating tropical ants and termites include the armadillo (*Dasypodidae*) and some pangolins (*Manis temminckii* and *M. gigantea*), tamanduas (*Myrmecophaga tridactyla*) and armadillos (*Prionomys maximus*). These^{4,5} and the sloth bear⁶ generally have lower standard rates of energy expenditure than other mammals. As ant and termite predators increase in size, their basal rate of metabolism decreases (Fig. 1), a trend that is particularly evident when species in a family are compared (to correct for any putative effect of phylogeny or ecological/behavioural uniformity).

A reduction in metabolic rate reduces the effective body size, which can be estimated from the total basal rate of the largest committed ant/termite eaters. If an all-mammal standard⁷ for basal metabolic rate

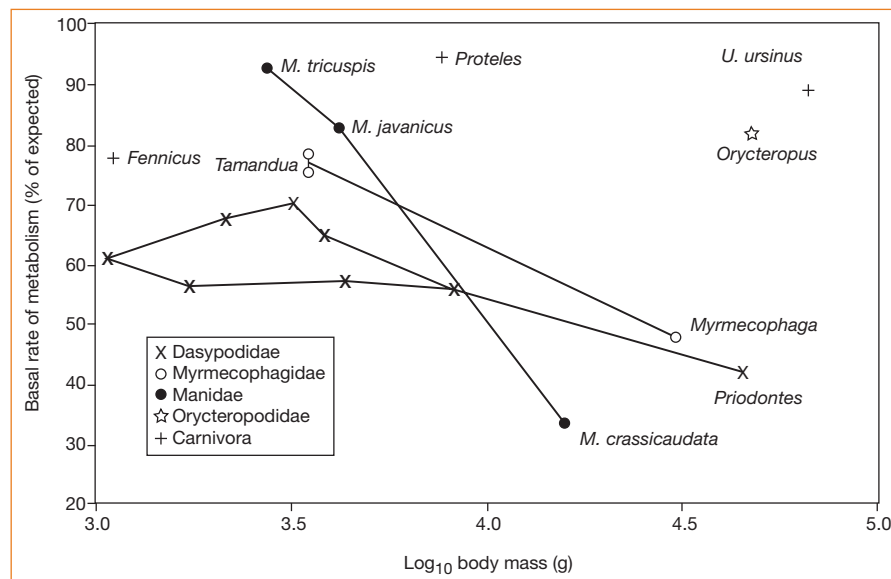


Figure 1 Basal rate of metabolism, expressed as a percentage of the basal rate expected from an all-mammal curve⁷, in various mammals^{4–6} that specialize on soil invertebrates, as a function of body mass. Species that belong to the same family are connected.

is used, a 15.9-kg *Manis crassicaudata* has the basal rate of a 3.4-kg standard mammal, a 30.6-kg *Myrmecophaga* has that of a 10.9-kg standard mammal, a 45.2-kg *Priodontes* has that of a 13.4-kg standard mammal, a 48-kg *Orycteropus* has that of a 36.2-kg standard mammal, and a 67.0-kg *U. ursinus* has that of a 56.5-kg standard mammal.

These calculations indicate that the maximum body mass in a standard mammal compatible with an ant/termite-eating habit is 11–13 kg, with the exception of the armadillo and sloth bear. This calculation may account for the comparatively high basal rate in *Proteles* (Fig. 1), which weighs less than 10 kg — at that mass, an adjustment of basal rate may not be required. What seems to be limited is the total rate of energy expenditure, not body mass: a limiting rate may be encountered in various masses at the expense of conforming to a standard curve and having effective endothermic temperature regulation.

Two of the species shown in Fig. 1 exceed the 11–13-kg limit to the ‘adjusted’ mass. The large mass and comparatively high basal rate of the sloth bear correlate with a diet that is about 50% fruit³, although it is not clear whether addition of fruit to the diet permits a higher expenditure or size. The most distinctive large terrestrial specialist insectivore is the armadillo, which conforms neither to the original analysis², nor to the evasion described here. How it can have its comparatively high basal rate and a large body mass, and eat only ants and termites, is unknown. Under the assumption that a limiting energy expenditure exists, some other evasion may apply.

A limit to the exclusive use of invertebrates by terrestrial mammals, if one exists, may be associated with the cost of prey col-

lection, which is why the largest species are tropical and feed on ants and termites: only these prey occur in sufficiently large colonies to make prey acquisition energetically feasible, and such large colonies occur only in the lowland tropics. In the absence of colonial ants and termites, terrestrial invertebrate-eaters might attain a maximal mass of 10 kg (ref. 5). The absence of an ant/termite specialization in large carnivores may occur because this niche was occupied by other mammals before the evolution or arrival of carnivores, the only opportunity available being at intermediate masses, which was exploited in Africa by the armadillo and the bat-eared fox (*Otocyon megalotis*).

Although it might be argued that this analysis fails to take phylogenetic history into consideration, it has been pointed out⁸ that ‘corrections’ for proposed phylogeny erroneously assume the priority of phylogeny as a factor influencing phenotypic characters, thus ignoring the complex interactions among determining factors. The model of Carbone *et al.*² is ultimately called into question because it ignores the residual variation and therefore the biological flexibility inherent in all scaling functions.

Brian K. McNab

Department of Zoology, University of Florida,
Gainesville, Florida 32611, USA
e-mail: mcnab@zoo.ufl.edu

1. McNab, B. K. *Funct. Ecol.* **6**, 672–679 (1992).
2. Carbone, C., Mace, G. M., Roberts, S. C. & Macdonald, D. W. *Nature* **402**, 286–288 (1999).
3. Laurie, A. & Seidensticker, J. *J. Zool. Lond.* **182**, 187–204 (1977).
4. McNab, B. K. *J. Mammal.* **61**, 606–627 (1980).
5. McNab, B. K. *J. Zool. Lond.* **203**, 485–510 (1984).
6. McNab, B. K. *J. Mammal.* **73**, 168–172 (1992).
7. McNab, B. K. *Q. Rev. Biol.* **63**, 25–54 (1988).
8. Westoby, M., Leishman, M. R. & Lord, J. M. *J. Ecol.* **83**, 531–534 (1995).

Phytochromes and light signal perception by plants—an emerging synthesis

Harry Smith

Department of Biology, University of Leicester, Leicester LE1 7RH, UK

For plants, the sensing of light in the environment is as important as vision is for animals. Fluctuations in light can be crucial to competition and survival. One way plants sense light is through the phytochromes, a small family of diverse photochromic protein photoreceptors whose origins have been traced to the photosynthetic prokaryotes. During their evolution, the phytochromes have acquired sophisticated mechanisms to monitor light. Recent advances in understanding the molecular mechanisms of phytochromes and their significance to evolutionary biology make possible an interim synthesis of this rapidly advancing branch of photobiology.

The phytochromes are signal-transducing photoreceptors that convert between inactive and active forms in response to different wavelengths of light. This conversion is used to synchronize plant development to the exigencies of the light environment. Light signals provide information of crucial ecological value at many stages in development. Seed germination, seedling establishment, the proper development of photosynthetic machinery, the architecture of the vegetative plant, the timing of flowering, tuberization and bud dormancy, the responses to neighbour competition, and the allocation of resources to root, stem, leaf, reproductive or storage structures are all potentially controlled by the perception of environmental light signals by the phytochromes. We can now begin to trace the biological significance of photoperception from the molecular mechanisms to the life of plants in their natural environment. Current research stretches from the photophysics of the light sensor molecules to the population biology of plants in the field. This article attempts an undoubtedly premature synthesis of a research area that is advancing at remarkable speed.

Phytochromes as sensors of space and time

Light signal perception may be thought of as the means by which the plant determines where it is in space and time. Although this sounds anthropomorphic, it nevertheless describes the significance of the perception of light signals by the phytochromes and other signal-transducing photoreceptors. The three classes of plant signal-transducing photoreceptors—phytochromes¹, cryptochromes² and phototropin³—have well characterized molecular details and functions. Each group of photoreceptors operates both independently and in concert with the others to regulate plant development. The ecophysiological functions of the phytochromes are determined by their photosensory characteristics, which in turn depend on photochemistry (see Box 1). The striking characteristic of the phytochromes is their reversible photochromism—the property of changing colour on photon absorption and of reverting to the original form on the absorption of another photon. The absorption maximum of the phytochrome 'Pr' form is close to that of the chlorophylls (red light), but the 'Pfr' form absorbs at a longer wavelength (far-red light). In effect, this means the phytochromes can be used as sensitive estimators of the spectral changes that happen within plant communities when daylight interacts with photosynthetic structures⁴. Daylight contains roughly equal proportions of red and far-red light (red:far-red ≈ 1.2), but within vegetation that ratio is lowered by the absorption of red light by photosynthetic pigments. Changes in the red:far-red ratio are much more reliable indicators of the proximity of potentially competing neighbours than the concomitant reductions in the total

amount of light penetrating the canopy. Within dense canopies, the far-red radiation scattered or reflected from leaves is a unique and unambiguous signal of the proximity of neighbours⁵, and hence of community density⁶. Plants use the phytochromes as proximity sensors and modify their growth and development, constituting the 'shade-avoidance syndrome'⁷. Upon sensing a low red:far-red ratio, a shade-avoiding plant will exhibit enhanced elongation growth and, if the stratagem is successful, will project its leaves into regions of unattenuated daylight. If elongation is unsuccessful, other aspects of the shade-avoidance syndrome cause accelerated flowering and early production of seeds, enhancing the probability of survival. Shade avoidance is a strategy employed by the majority of angiosperms, ranging from small herbs to large trees, and is of major ecological advantage. Phytochrome-mediated proximity sensing, in effect, provides the plant with positional information with respect to potential competing neighbours.

Phytochromes also provide plants with temporal signals that entrain the phases of the biological clock, and others that ensure crucial developmental steps are initiated at appropriate points in the life cycle. Endogenous circadian rhythms synchronize development to the changing seasons, as exemplified in the photoperiodic control of flowering and dormancy. Even when employed as simple light detectors—such as in the stimulation of seed germination or the conversion of the etiolated seedling to photosynthetic competence—the phytochromes may be thought of as timing agents. Germination of the seed and maturation of the developing seedling, both dependent upon limited stored reserves, are probably the most vulnerable stages of the life cycle, and getting the timing right must be vital to long-term survival. In these processes, the phytochromes do not work alone; the cryptochromes are often responsible for initiating germination and they have important roles in de-etiolation. One implication of this may be that getting the timing right is so crucial to survival that reliance on one set of environmental light signals alone is insufficient to guarantee success under all conditions.

Evolution of phytochrome function

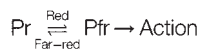
The phytochrome genes encode a small family of photoreceptors. In the model plant *Arabidopsis thaliana*, there are five members—phytochrome A (phyA) to phytochrome E (phyE)⁸. (Nomenclature usage differs in some species.) Sequence comparisons from a wide range of angiosperms indicate that the maximum family size is likely to be five, although monocotyledonous plants may not contain phyE homologues⁹. Phylogenetic analyses of higher plant *PHY* sequences point to four major gene duplication events in the evolution of the *PHY* genes¹⁰. The first, occurring about the time of the origin of seed plants, generated the *PHYA/C* and *PHYB/D/E* lines. Two later dupli-

Box 1

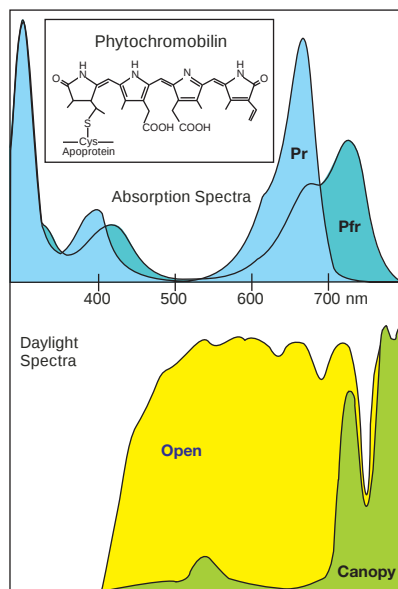
Sensors of environmental light signals

Phytochromes are photochromic photoreceptors

The central hypothesis of phytochrome action, proposed almost 50 years ago from pioneering investigations by S. Hendricks, H. Borthwick and colleagues, is that the photoreceptors exist in two, photoconvertible forms, Pr and Pfr. Pr is biologically inactive and upon absorption of red photons is converted to Pfr, the active form. Pfr is converted back to Pr by far-red photons. Biological action stems from Pfr.



The photoconversions involve a number of intermediate forms in both directions, and the establishment of an equilibrium between Pr and Pfr takes several minutes even at daylight irradiance levels. The absorption spectra of the phytochromes peak at about 665 nm and 730 nm. The absorption bands overlap, so radiation below about 700 nm activates photoconversion of both Pr and Pfr. Thus, in daylight for example, a photoequilibrium of about 60% Pfr/P (where P = total phytochrome) is established—in canopy shade or crowded communities, the photoequilibrium can be as low as Pfr/P = 0.1. This is the basis of the shade-avoidance syndrome.



Phytochrome absorption spectra drawn from data supplied by J. C. Lagarias.

Chromophore is a linear tetrapyrrole

The chromophore is an open-chain linear tetrapyrrole—known as phytochromobilin—and is closely related to phycocyanobilin, the chromophore of the abundant algal pigment C-phycocyanin. In higher plant phytochromes, the chromophore is covalently attached to the protein through a thio-ether link at a cysteine positioned at amino-acid residue 374 (numbering for phyA). Assembly of apoproteins and chromophore occurs spontaneously, presumably involving inherent chromophore lyase activity in the phytochrome apoproteins. This property has allowed the construction of recombinant phytochrome adducts with either phytochromobilin or phycocyanobilin—both are spectrally photoreversible and active when transgenically expressed *in planta*.

cations, at about the time of the origin of the flowering plants, separated *PHYA* from *PHYC* and *PHYB/D* from *PHYE*. *PHYB* and *PHYD* diverged much more recently. Evolution within the phytochrome family appears to be faster than other plant nuclear genes¹¹.

These evolutionary pathways created a family of proteins that detect identical environmental signals but employ those signals in different functions. *PHY*-like genes occur not only in all green plants, including gymnosperms, ferns, mosses and algae, but also in cyanobacteria^{12–14} and even in certain other bacteria^{15,16}. The chromophore-bearing domain is conserved between plants and bacteria, and in most cases the bacterial chromophore is a bilin, as it is in plants. In the purple photosynthetic bacterium *Rhodospirillum rubrum*, the bilin attachment site is unoccupied, but the domain is linked to a photoactive yellow protein (PYP) domain carrying *p*-hydroxycinnamic acid as a chromophore¹⁵. In the non-photosynthetic bacteria *Deinococcus radiodurans* and *Pseudomonas aeruginosa*, the chromophore is a bilin, but its attachment is through a Schiff's base to a histidine, rather than a thio-ether linkage to the cysteine attachment site present in plant phytochromes¹⁶. Thus, although the chromophore is not always a bilin or attached to the same amino-acid residue, the conserved region seems to be involved in photoperception in all organisms investigated. This might provide a clue to the evolutionary origin of the conserved chromophore-bearing domain of the phytochromes.

In the mosses and ferns, the phytochromes seem to be particularly involved in phototropism¹⁷, a function mediated exclusively by the blue-light absorbing phototropin in the angiosperms³. Indeed, in the fern *Adiantum*, a gene has been characterized that encodes both a typical phytochrome and a protein with sequence similarity to NPH1 (the *Arabidopsis* phototropin)¹⁸. The evolution of a chimaeric photoreceptor bearing properties of both red/far-red and blue light absorption poses the question whether such molecules exist and function elsewhere. Another chimaeric phytochrome gene has been identified in the moss *Ceratodon*¹⁹, encoding a protein kinase carboxy-terminal segment, but whether this is functional has not been determined.

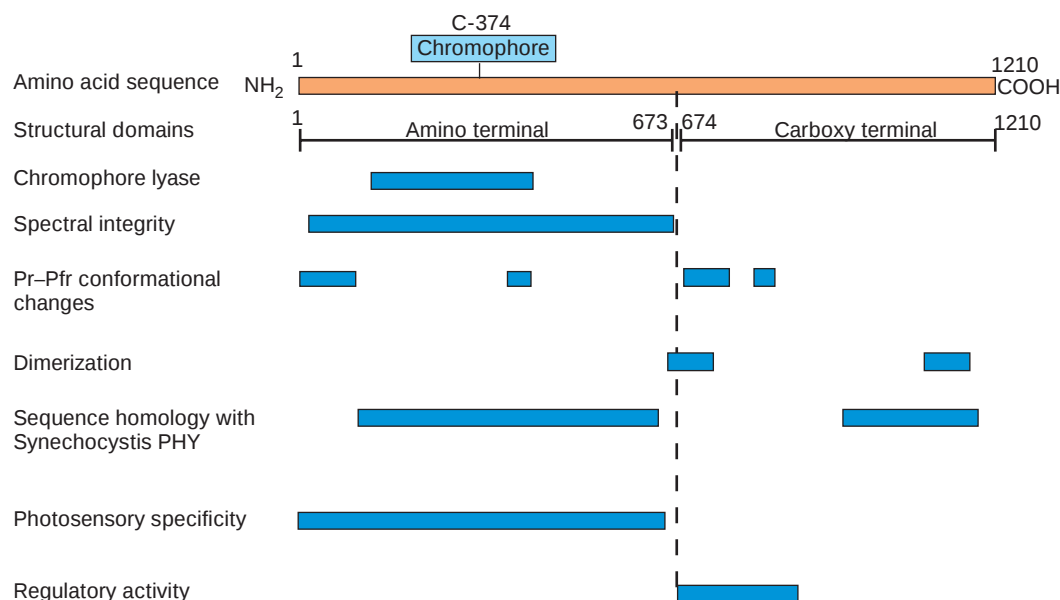
In the angiosperms, the phytochromes have acquired new functions and fine-tuned existing functions during evolution (Box 2). The individual family members have both distinct and overlapping physiological functions²⁰, and even antagonistic interactions have been reported²¹. It is important to realize that phytochromes operate quantitatively to limit the rates of ongoing processes, rather than qualitatively to select between potential developmental fates. In other words, phytochromes are more like speed limits than signposts on the highways of development.

Phytochrome mechanisms

Most investigators propose that the phytochromes act through the selective regulation of gene expression. The expression of several genes, largely those encoding enzymes and other components of the photosynthetic machinery, is regulated by light, often by phytochrome^{22,23}. On the other hand, some phytochrome-mediated responses are best explained by assuming rapid alterations in inter- or intra-cellular ion balances²⁴. Some growth responses also occur quite rapidly (within minutes^{25,26}), posing the question whether transduction chains involving the regulation of nuclear gene expression would be sufficiently quick. In certain algae, mosses and ferns, at least part of the cellular phytochrome is cytoplasmic and dichroically organized with respect to either the cell membrane or the cytoskeleton²⁷. These phenomena imply a cytoplasmic growth-homeostasis mechanism in which the photoreceptor continuously monitors the light environment and mediates rapid changes in ionic concentrations within cell compartments that modulate extension growth. Therefore, individual phytochromes may have at least two, separate mechanisms of action: one that results in selective expression of target genes and another that rapidly and reversibly operates to modulate cellular ionic balances.

Box 2

Molecular, physiological and ecological functions



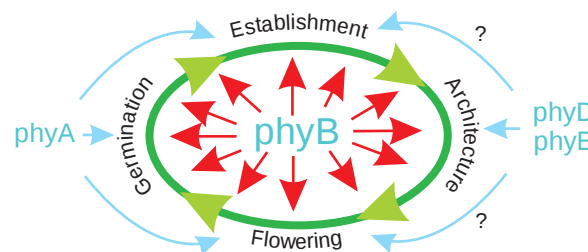
Redrawn and simplified with permission from ref. 69.

The phytochromes are chromoproteins with a relative molecular mass of about 120,000–126,000. They are encoded by five *PHY* genes (named *PHYA–PHYE*) in the model crucifer *Arabidopsis thaliana* (family size and nomenclature may vary in other species). Isolation of mis-sense mutations and transgenic expression of mutant phytochromes enabled certain molecular functions to be attributed to 'domains'. Thus, the molecule possesses two structural domains: a globular N-terminal half and a more linear C-terminal half. Chromophore lyase and spectral integrity are associated with the chromophore-bearing N-terminal domain. The phytochromes are dimers in solution, with putative dimerization sites located within the C-terminal segment. Evidence from photophysical analysis indicates that conversion from Pr to Pfr causes an 'opening' of the protein conformation in relation to the chromophore, perhaps facilitating the interaction of Pfr with its primary reaction partners. Transgenic expression of constructs with the N- and C-terminal domains exchanged between *phyA* and *phyB* show the photosensory specificity is in the N-terminal half, and the regulatory specificity is in the C-terminal half. Alignment with gene sequences from cyanobacteria reveals common ancestry and provokes interesting questions on the original adaptive value of the phytochromes.

Physiological and ecological functions

It is convenient to think of the phytochromes as being 'single-input/multiple-output' sensory systems, in which a common signal evokes a range of biological responses. Phytochromes are capable of regulating almost all phases of plant development, but the control is conditional or facultative, rather than obligatory.

Germination. Many small seeds with low levels of stored reserves require light signals for germination, whereas most larger seeds do not. For some buried seeds, the sensitivity is spectacular, sometimes requiring only milliseconds of light exposure for full response.



Seedling establishment. This is the set of processes through which a germinated seedling becomes photoautotrophic. Phytochromes co-operate with cryptochromes to regulate de-etiolation, leaf expansion and chloroplast maturation.

Architecture of the mature plant. The size and disposition of internodes and leaves, the balance between main stem and lateral branches, the angles of petioles and leaf laminae are acutely sensitive to the radiation environment.

Induction of flowering and dormancy. The photoperiodic perception of day length, involving interaction with the biological clock, is necessary (in some plants) for the induction of flowering and bud dormancy.

Towards a life-history of phytochrome functions

Mutant studies⁷⁰ and the use of transgenic plants expressing individual phytochrome genes⁷¹ mean that we can reliably allocate physiological and ecological function to four of the five phytochromes, at least for *Arabidopsis thaliana*. Mutants null for phytochromes A, B, D and E have been isolated and their physiological responses extensively characterized. The overall picture is that *phyB* has a role at all stages of the life cycle, whereas *phyA*, *phyD* and *phyE* exert their principal functions at selected stages.

Phytochrome action requires two partial processes at the level of the photoreceptor molecule: perception of the light signal and its transduction to a biochemical signal. Domain-swapping experiments in which the amino-terminal and the carboxy-terminal halves of phyA and phyB were exchanged verify the molecular specificity of these sensory and regulatory activities with the photosensory functions of phyA and phyB present in the respective N-terminal half of the molecule, and a common regulatory site present in the C-terminal portion²⁸. Thus, phytochromes exhibit dual molecular functions: a sensory function responsible for detecting relevant light signals, and a regulatory function in which the perceived information is transferred to downstream transduction pathways. Interestingly, the sensory and regulatory domains appear to be evolving at different rates, with the C-terminal domain evolving at least twice as fast as the N-terminal domain¹¹.

Current ideas on the primary mechanism of phytochrome regulation of gene expression centre on two contrasting, but not necessarily incompatible, hypotheses (Fig. 1). In one, phytochromes are considered to be kinases that act on multiple substrates thereby regulating the expression of genes differentially. The other is that phytochromes have one or more specific reaction partners that direct signal transduction towards the selective control of gene expression.

Phytochromes as kinases. The possibility that plant phytochromes are kinases has long been controversial²⁹. In the prokaryotes, the *PHY* genes encode the sensory halves of two-component response systems, and significant sequence similarity exists between the C-terminal domains of eukaryotic phytochromes and bacterial response elements³⁰. The bacteriophytochrome (named *DrBphP*) in *D. radiodurans* functions as a light-regulated histidine kinase, controlling the synthesis of the carotenoid pigment deinoxanthin, which apparently serves to protect the organism from bright light¹⁶. Immediately downstream of *BphP* is a coding region for a response regulator protein (*BphR*), which acts as the phosphate acceptor for the kinase activity of *DrBphP*. The closest sequence relative of *BphR* is the corresponding response regulator component of the *Synechocystis* phytochrome (*Rcp1*), the proposed phosphate acceptor for the cyanobacterial phytochrome, *Cph1* (ref. 14). The kinase activity of plant phytochromes was confirmed by the construction of recombinant phytochromes using *PHY* genes from *Arabidopsis* and the green alga *Mougeotia*. After assembly with chromophore, the holoproteins possessed serine/threonine kinase activity, rather than the expected histidine kinase activity³¹. The recombinant phytochromes autophosphorylated and, in a clinching experiment, transferred phosphate to recombinant *Rcp1*.

Support for the phytochrome-as-kinase concept comes from the detection, by yeast two-hybrid screening, of a phytochrome kinase substrate (PKS1), a cytosolic protein that accepts a phosphate from phyA³². Phosphorylation is on a serine, and to a lower extent on a threonine residue and is regulated by light, being about twofold higher with Pfr than Pr. Phosphorylation also occurs *in vivo* in a phytochrome-dependent manner. Transgenic expression shows that PKS1 is a negative regulator of photomorphogenesis specific to phyB. Other possible substrates of phytochrome kinase activity are the cryptochrome photoreceptors³³. The situation is even more complex as phyA also interacts with a nucleoside diphosphate kinase (NDPK2), found in both the cytoplasm and the nucleus³⁴. The kinase activity of NDPK2 is increased about twofold when bound to recombinant phyA in the Pfr form, but an *in vitro* activation of NDPK1 by red or far-red radiation has not been reported. The regulation of gene expression could, in principle, emanate from the kinase activity of phytochrome *per se*, and/or activation of NDPK1.

Primary interaction partners. Yeast two-hybrid screens have revealed several potential primary reaction partners for phyA and phyB. Whether all of these will have important roles is in the balance, but the most conclusive evidence yet is for PIF3 (phytochrome interacting factor 3; ref. 35). PIF3 was found in a two-hybrid assay using the C-terminal portion of phyA as the bait. It interacts equally with the C-

terminus of phyA and phyB, but preferentially with the intact phyB protein. Interaction with intact phyB is light dependent³⁶. PIF3 is a basic helix—loop—helix (bHLH) protein with a PAS motif and locates to the nucleus in transfection experiments. PIF3 binds poorly to mis-sense phyB molecules mutated in the putative C-terminal regulatory region, and furthermore, transgenic expression of sense and anti-sense constructs of PIF3 perturbs the response to light. Serendipitously, PIF3 was simultaneously identified as a transduction chain component by a screen for gain-of-function mutants under red radiation treatment³⁷. These data strongly support the hypothesis that PIF3 is a functional primary reaction partner for phyB.

Recent data on the phyB/PIF3 interaction point to a remarkable shortcut mechanism in which light signals are targeted directly through phyB to PIF3 bound to promoter elements of some light-regulated genes³⁸. Promoter analysis has identified a number of *cis*-acting light-responsive elements (LREs) and some DNA-binding

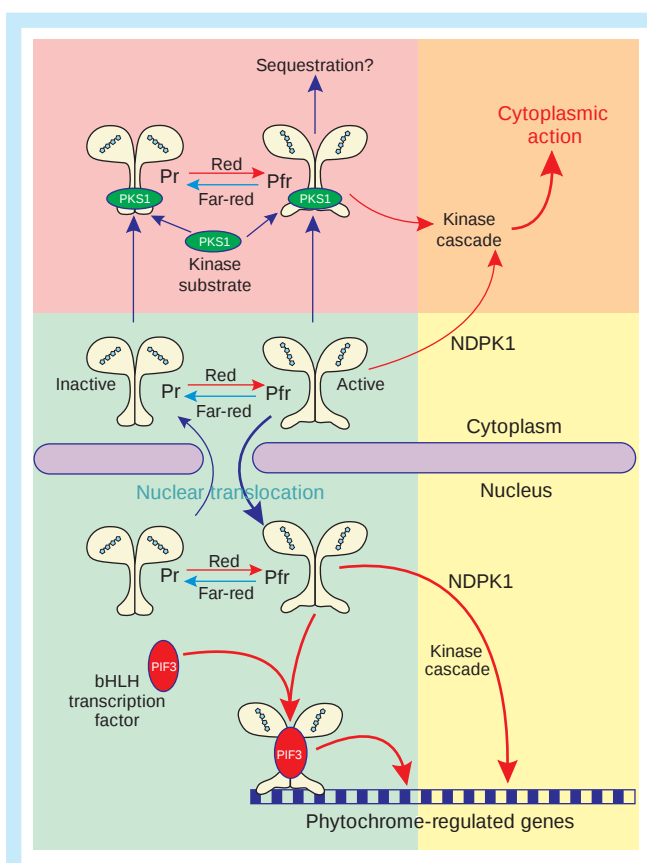


Figure 1 Diagram of phytochrome action. Phytochromes undergo photoconversion from the biologically inactive form (Pr) to the active form (Pfr). Pr and Pfr are shown as dimers in the cell. The Pr–Pfr conversions are initiated by photon absorption in the chromophore leading to steric changes, causing the holoprotein to ‘open up’ and facilitating interaction with putative reaction partners. The diagram shows the three major theories for the subsequent actions of the phytochromes, although Pfr may regulate growth and development by other processes. Pink area: both Pr and Pfr interact with PKS1, the phytochrome kinase substrate, in the cytosol. This may be the first step in a kinase cascade (orange area) culminating in action within the cytoplasm. Alternatively, interaction with PKS1 may result in sequestration of phytochrome in the cytosol, preventing translocation to the nucleus. Yellow area: Pfr interacts with NDPK1, a nucleoside diphosphate kinase, which is located both in the cytoplasm and the nucleus. Again, this interaction may initiate a kinase cascade (orange) leading to ultimate action within the cytoplasm and/or nucleus. Green area: Pfr translocates to the nucleus and Pfr is translocated back to the cytoplasm. The weights of the arrow emphasize the differential rates of import and export. Within the nucleus, Pfr binds with PIF3 (phytochrome interacting factor 3) which is located exclusively within the nucleus. PIF3 is a basic helix–loop–helix transcription factor that binds to the promoters of selected light-regulated genes in combination with Pfr and initiates or enhances transcription.

proteins that may interact with LREs and thus regulate transcription³⁹. One of these promoter elements, a G-box motif, is the core PIF3 target element³⁸. So, PIF3 associates with elements characteristically found in the promoters of certain genes whose expression is regulated by phytochrome. Moreover, phyB in the Pfr form complexes with PIF3 bound to its DNA target site. PIF3 anti-sense plants displayed reduced levels of phytochrome-regulated gene expression for some, but not all, light-regulated genes. These data provide a picture of the direct transfer of light signals to the promoters of a subset of light-regulated genes.

Light-regulated nuclear translocation

The phytochrome apoproteins are synthesized within the cytosol and assemble autocatalytically with the plastid-derived chromophore. For years phytochromes were considered to be entirely cytosolic, but now there is strong evidence for photoactivated nuclear translocation of phytochromes⁴⁰. Both phyA and phyB tagged with green fluorescent protein (GFP) show light-activated import into the nuclei of tobacco⁴¹ and *Arabidopsis*⁴² cells. Import of phyB occurs only in the Pfr form and is slow, requiring several hours for full mobilization. Phytochrome A, either as Pfr_A, or as Pr_A that has been photoconverted through Pfr_A and back again, moves more rapidly (about 15 min). The photobiological criteria are satisfied as phyB transport is activated by red radiation and inhibited by far-red radiation, whereas the transport of phyA is maximal under continuous far-red radiation. These facts provide a framework for understanding the phytochrome regulation of gene expression through the translocation of Pfr into the nucleus, interaction with primary reaction partners (such as PIF3) and direct regulation of the promoters of light-regulated genes. Alternatively, nuclear localization places Pfr in the appropriate cellular compartment for its activity as a protein kinase to operate on factors regulating transcription.

It is worth considering whether all the correct questions are being asked. The data on nuclear transport and gene expression described come from experiments on dark-grown seedlings given light treatments. Seedlings treated in this way are induced to undergo a one-off transition, which will not be repeated, and which under natural conditions prepares the plant for life in a photic environment. The dynamics of nuclear translocation may only be mechanistically significant during this initial transition. Thereafter, except for newly synthesized molecules, the phytochrome may remain principally within the nucleus during daylight hours, where interaction with light-responsive genes can occur rapidly. Superimposed on this there will be daily gross movements of phytochrome into and out of the nucleus conditioned by light and dark periods, and indeed it has already been reported that nuclear translocation of phyB is under circadian control⁴³. Such control may position phyB in the nucleus, where it can rapidly control development in response to signals from neighbouring plants, and so provide the sensitivity to environmental fluctuations that is characteristic of shade avoidance in plant communities.

Phenotypic plasticity and evolutionary significance

Phytochrome-mediated shade avoidance is a model for the functional significance of physiological adaptation to environmental signals, giving ecologists and evolutionary biologists a handle on the evolution of phenotypic plasticity. Phenotypic plasticity—the expression of variability in the phenotype of individuals of identical genotype—and its fitness consequences are crucial concerns for evolutionary biologists^{44–46}. There is controversy regarding the genetic mechanisms for the evolution of plasticity^{47,48}, and clear-cut experimental approaches are necessary to examine whether plasticity is adaptive, to what extent it is constrained by trade-offs and linkage restrictions, and how plasticity evolves. Phytochrome-mediated shade avoidance is a successful example of two-way exchange between the reductionism of physiological and molecular analysis and the holistic approach of evolutionary biology⁴⁹.

The unique and unambiguous environmental signal of far-red

radiation scattered from neighbouring vegetation evokes a range of plastic outputs—enhanced elongation, strengthened apical dominance, elevated leaf angle, altered resource allocation and accelerated flowering—that are assumed to allow adaptation to competition from those neighbours. Shade avoidance therefore represents a classic single-input, multiple-output system. The range of plasticity is huge, and the assumption that such plasticity is adaptive is understandable, but it has rarely been explicitly tested. Indeed it is difficult to test because the plasticity itself prevents the expression of inappropriate phenotypes in any given environment⁵⁰. Though crude, some approaches are possible, such as using mutant^{51,52} or transgenic⁵³ plants disabled in signal perception, and these have supported the adaptive plasticity hypothesis. Plasticity observed as variation in physiological output must result from the selective expression of genetic variation within these transduction pathways⁵⁴. To extend the earlier metaphor, as an individual plant responds to fluctuating environmental light signals, the speed limits on the branched developmental highways must be variously imposed or relieved to provide appropriate responses.

What is not so obvious is that micro-evolution may operate differentially on the pathways emanating from the same signal. In a study of more than 100 *Arabidopsis* ecotypes (accessions) with wide variations in response to reflected far-red light, there was no correlation between hypocotyl elongation and accelerated flowering. That is, ecotypes that responded strongly in elongation did not necessarily respond strongly in floral acceleration, and *vice versa* (J. Botto and H.S., unpublished data). When searching for the molecular basis of such adaptive diversity within a single species, several targets of interest emerge. First, molecular variation in the photoreceptor genes may lead to differential transduction of reflection signals. Second, downstream components specific to photomorphogenesis may vary between ecotypes. Third, regulatory genes located in the basic ground plan of development, but whose activity comes under the influence of the phytochromes, may have been under selection pressure that resulted in differential output from the common signal. In this regard, regulatory genes (that is, loci that control the expression of other genes) are increasingly interesting to evolutionary biologists. Such genes are expected to be important in the genetic architecture of development, and functional polymorphism among such genes within species is likely to contribute significantly to micro-evolution^{55,56}. Regulatory genes encode sequence-specific DNA-binding transcriptional activators or cell–cell signalling molecules⁵⁷, and some evolve more rapidly than structural genes^{58,59}. To search for the basis of rapidly evolving differences in phenotypic plasticity to the light environment, the best way may be to concentrate on potential regulatory genes specific to phytochrome-mediated signal transduction.

The *PHY* genes are evolving rapidly¹¹, but they do not fall into the strict definition of regulatory genes, even though their gene products do regulate the expression of other genes through transcription factors such as PIF3. Molecular polymorphism of phytochromes within species probably exists, and may result in different responsiveness to light signals in different genotypes. However, this would not account for situations where a single genotype has different responsiveness to a single signal. Downstream components fall more readily into the regulatory gene category than the sensor genes themselves, as some, at least, operate through the regulation of expression of other genes. Therefore important information may be gained by a search for functional sequence diversity in putative regulatory genes within the phytochrome transduction pathways or for regulatory genes whose expression varies specifically in relation to phytochrome-mediated signal transduction.

The level of quantitative complexity is daunting, with many large gene families potentially displaying high within-species variation in expression, temporally and spatially, as part of the developmental changes elicited by the reflection signal. Genomic and proteomic technology will allow the analysis of molecular variation, but to identify the crucial elements that provide adaptive advantage, genetic

analysis combined with ecological assessment of fitness attributes will be necessary. Recently, great improvements have been made in the application of quantitative genetic analysis to problems of plant developmental physiology⁶⁰ and in the exploitation of natural genetic variation⁶¹. Important quantitative trait loci (QTL) are being rapidly identified for, *inter alia*, seed size⁶², flowering time⁶³ and circadian rhythms⁶⁴, and, coupled with molecular marker technology, will in principle allow the characterization of the genes themselves. Application of these approaches to phytochrome-mediated phenotypic plasticity may yield the most important information. Having located QTL specific for phytochrome-mediated responses of ecological significance, studies of QTL nucleotide diversity and its relationship to phenotype should provide a basis for a comprehensive understanding of the genetic architecture of photomorphogenesis.

Prospect and speculation

This field is moving so rapidly that this review will already be less than current when published. We are likely to see imminent advances in the understanding of the molecular and cellular mechanisms of the photoregulation of gene expression, although the equally important mechanisms that modulate cell growth rates have not generated a similar intensity of interest. The application of genetic analysis to the evolutionary and ecological questions raised here is likely to be productive in the medium term, although it will initially be restricted to model plants such as *Arabidopsis*. There is no doubt that exploitation of photomorphogenesis to improve crop plant growth will generate many new approaches. Already, manipulation of the expression of phyA has led to major changes in plant architecture in the field, with predictions of substantial increases in harvestable yield per plant⁶⁵. There are dozens of laboratories world-wide working on the development of genetically modified crop plants using the phytochrome genes. Other applications include those that modify plant responses to photoperiod, which may be of great benefit for certain crops⁶⁶.

To end with a speculation, it seems particularly fortuitous that the phyB/D/E branch of the phytochrome family evolved from the phyC/A branch coincidentally with the evolution of the seed plants, and that its further subdivision coincided with, or preceded, the evolution of the flowering plants¹⁰. Phytochromes B, D and E act singly, collectively, and to some extent redundantly, to mediate proximity perception and shade avoidance. Gene duplication and function acquisition within this phytochrome subfamily has reached a stage where the perception of the relative proportions of red and far-red radiation within plant communities provides an exceptionally sensitive response to potential neighbour competition. This capacity is most highly expressed, and most widespread, in the angiosperms. Ferns and mosses generally react to density by shade tolerance reactions, and of non-angiosperm plants, only the gymnosperms seem capable of true shade avoidance, and that is limited in extent and distribution. Perhaps the evolution of the ability to detect light signals reflected from neighbours was crucial to the advance of the angiosperms to their present dominant status within the plant kingdom—without phytochrome B, would we still be surrounded by a carboniferous flora? This notion is far-fetched and tongue-in-cheek, but it could be supported by studies using gene duplication analysis within the phytochromes as an approach toward rooting the angiosperms^{67,68}. Phytochromes originated in prokaryotic progenitors of present-day plants, presumably acting as simple light sensors. Their unique capacity for photochromic conversion between two molecular forms may have been functionally insignificant within the early prokaryotes, but this property has been selected and refined through the evolution of the land plants into a multi-component sensor system of equal global significance to that of vision in animals. □

- Neff, M. M., Fankhauser, C. & Chory, J. Light: an indicator of time and place. *Genes Dev.* 14, 257–271 (2000).
- Ahmad, M. & Cashmore, A. R. Seeing blue: The discovery of cryptochrome. *Plant Mol. Biol.* 30, 851–861 (1996).

- Christie, J. M. *et al.* *Arabidopsis* NPH1: A flavoprotein with the properties of a photoreceptor for phototropism. *Science* 282, 1698–1701 (1998).
- Smith, H. Light quality, photoperception and plant strategy. *Annu. Rev. Plant Physiol.* 33, 481–518 (1982).
- Ballare, C. L., Scope, A. L., Sanchez, R. A., Casal, J. J. & Ghersa, C. M. Early detection of neighbour plants by phytochrome perception of spectral changes in reflected sunlight. *Plant Cell Env.* 10, 551–557 (1987).
- Gilbert, I. R., Seavers, G. P., Jarvis, P. G. & Smith, H. Photomorphogenesis and canopy dynamics. Phytochrome-mediated proximity perception accounts for the growth dynamics of canopies of *Populus trichocarpa* X *deltoides* 'Beaupré'. *Plant Cell Env.* 18, 475–497 (1995).
- Smith, H. Physiological and ecological function within the phytochrome family. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 46, 289–315 (1995).
- Clack, T., Mathews, S. & Sharrock, R. A. The phytochrome apoprotein family in *Arabidopsis* is encoded by five genes: the sequences and expression of *PHYD* and *PHYE*. *Plant Mol. Biol.* 25, 413–427 (1994).
- Mathews, S. & Sharrock, R. A. The phytochrome gene family in grasses (Poaceae): A phylogeny and evidence that grasses have a subset of the loci found in dicot angiosperms. *Mol. Biol. Evol.* 13, 1141–1150 (1996).
- Mathews, S. & Sharrock, R. A. Phytochrome gene diversity. *Plant Cell Env.* 20, 666–671 (1997).
- Alba, R., Kelmenson, P. M., Cordonnier-Pratt, M. M. & Pratt, L. H. The phytochrome gene family in tomato and the rapid differential evolution of this family in angiosperms. *Mol. Biol. Evol.* 17, 362–373 (2000).
- Kehoe, D. M. & Grossman, A. R. Similarity of a chromatic adaptation sensor to phytochrome and ethylene receptors. *Science* 273, 1409–1412 (1996).
- Hughes, J. *et al.* A prokaryotic phytochrome. *Nature* 386, 663 (1997).
- Yeh, K. C., Wu, S. H., Murphy, J. T. & Lagarias, J. C. A cyanobacterial phytochrome two-component light sensory system. *Science* 277, 1505–1508 (1997).
- Jiang, Z. Y. *et al.* Bacterial photoreceptor with similarity to photoactive yellow protein and plant phytochromes. *Science* 285, 406–409 (1999).
- Davis, S. J., Vener, A. V. & Vierstra, R. D. Bacteriophytochromes: phytochrome-like photoreceptors from nonphotosynthetic eubacteria. *Science* 286, 2517–2520 (1999).
- Esch, H., Hartmann, E., Cove, D., Wada, M. & Lamparter, T. Phytochrome-controlled phototropism of protonemata of the moss *Ceratodon purpureus*: physiology of the wild type and class 2 pt-mutants. *Planta* 209, 290–298 (1999).
- Nozue, K. *et al.* A phytochrome from the fern *Adiantum* with features of the putative photoreceptor NPH1. *Proc. Natl Acad. Sci. USA* 95, 15826–15830 (1998).
- Thümler, F., Dufner, M., Kreisl, P. & Ditttrich, P. Molecular cloning of a novel phytochrome gene of the moss *Ceratodon purpureus* which encodes a putative light-regulated protein-kinase. *Plant Mol. Biol.* 20, 1003–1017 (1992).
- Whitelam, G. C. & Devlin, P. F. Roles of different phytochromes in *Arabidopsis* photomorphogenesis. *Plant Cell Env.* 20, 752–75 (1997).
- Smith, H., Xu, Y. & Quail, P. H. Antagonistic but complementary actions of phytochromes A and B allow optimum seedling de-etiolation. *Plant Physiol.* 114, 637–641 (1997).
- Tobin, E. M. & Kehoe, D. M. Phytochrome regulated gene expression. *Semin. Cell Biology* 5, 335–346 (1994).
- Kuno, N., Muramatsu, T., Hamazato, F. & Furuya, M. Identification by large-scale screening of phytochrome-regulated genes in etiolated seedlings of *Arabidopsis thaliana* using a fluorescent differential display technique. *Plant Physiol.* 122, 15–22 (2000).
- Shacklock, P. S., Read, N. D. & Trewavas, A. J. Cytosolic free calcium mediates red light-induced photomorphogenesis. *Nature* 358, 753–755 (1992).
- Smith, H., Jackson, G. M. Rapid phytochrome regulation of wheat seedling extension. *Plant Physiol.* 84, 1059–1062 (1987).
- Parks, B. M. & Spalding, E. P. Sequential and coordinated action of phytochromes A and B. *Proc. Natl Acad. Sci. USA* 96, 14142–14146 (1999).
- Wada, M., Grolig, F. & Haupt, W. Light-oriented chloroplast positioning—contribution to progress in photobiology. *J. Photochem. Photobiol. B* 17, 3–25 (1993).
- Quail, P. H. *et al.* Phytochromes—photosensory perception and signal-transduction. *Science* 268, 675–680 (1995).
- Cashmore, A. R. Higher-plant phytochrome: "I used to date histidine, but now I prefer serine". *Proc. Natl Acad. Sci. USA* 95, 13358–13360 (1998).
- Schneider-Poetsch, H. A. W. Signal transduction by phytochrome—phytochromes have a module related to the transmitter modules of bacterial sensor proteins. *Photochem. Photobiol.* 56, 839–846 (1992).
- Yeh, K. C. & Lagarias, J. C. Eukaryotic phytochromes: Light-regulated serine/threonine protein kinases with histidine kinase ancestry. *Proc. Natl Acad. Sci. USA* 95, 13976–13981 (1998).
- Fankhauser, C. *et al.* PKS1, a substrate phosphorylated by phytochrome that modulates light signaling in *Arabidopsis*. *Science* 284, 1539–1541 (1999).
- Ahmad, M., Jarillo, J. A., Smirnova, O. & Cashmore, A. R. The CRY1 blue light photoreceptor of *Arabidopsis* interacts with phytochrome A *in vitro*. *Mol. Cell* 1, 939–948 (1998).
- Choi, G. *et al.* Phytochrome signalling is mediated through nucleoside diphosphate kinase 2. *Nature* 401, 610–613 (1999).
- Ni, M., Tepperman, J. M. & Quail, P. H. PIF3, a phytochrome-interacting factor necessary for normal photoinduced signal transduction, is a novel basic helix-loop-helix protein. *Cell* 95, 667–677 (1998).
- Ni, M., Tepperman, J. M. & Quail, P. H. Binding of phytochrome B to its nuclear signalling partner PIF3 is reversibly induced by light. *Nature* 400, 781–784 (1999).
- Halliday, K. J., Hudson, M., Ni, M., Qin, M. M. & Quail, P. H. poc1: An *Arabidopsis* mutant perturbed in phytochrome signaling because of a T DNA insertion in the promoter of PIF3, a gene encoding a phytochrome-interacting bHLH protein. *Proc. Natl Acad. Sci. USA* 96, 5832–5837 (1999).
- Martínez-García, J. F., Huq, E. & Quail, P. H. Direct targeting of light signals to a promoter element-bound transcription factor. *Science* 288, 859–863 (2000).
- Terzaghi, W. B. & Cashmore, A. R. Light-regulated transcription. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 46, 445–474 (1995).
- Nagy, F. & Schäfer, E. Nuclear and cytosolic events of light-induced, phytochrome-regulated signaling in higher plants. *EMBO J.* 19, 157–163 (2000).
- Kircher, S. *et al.* Light quality-dependent nuclear import of the plant photoreceptors phytochromes A and B. *Plant Cell* 11, 1445–1456 (1999).

42. Yamaguchi, R., Nakamura, M., Mochizuki, N., Kay, S. A. & Nagatani, A. Light-dependent translocation of a phytochrome B-GFP fusion protein to the nucleus in transgenic *Arabidopsis*. *J. Cell Biol.* **145**, 437–445 (1999).
43. Bognar, L. K. *et al.* The circadian clock controls the expression pattern of the circadian input photoreceptor, phytochrome B. *Proc. Natl Acad. Sci. USA* **96**, 14652–14657 (1999).
44. Bradshaw, A. D. Evolutionary significance of phenotypic plasticity in plants. *Annu. Rev. Genet.* **13**, 115–155 (1965).
45. Schlichting, C. D. The evolution of phenotypic plasticity in plants. *Annu. Rev. Ecol. Systematics* **17**, 667–693 (1986).
46. Sultan, S. E. Evolutionary implications of phenotypic plasticity in plants. *Evol. Biol.* **21**, 127–178 (1987).
47. Via, S. *et al.* Adaptive phenotypic plasticity: consensus and controversy. *Trends Ecol. Evol.* **10**, 212–216 (1995).
48. Van Tienderen, P. H. & Koelwijn, H. P. Selection on reaction norms, genetic correlations and constraints. *Genet. Res.* **64**, 115–125 (1994).
49. Callahan, H. S., Pigliucci, M. & Schlichting, C. D. Developmental phenotypic plasticity: where ecology and evolution meet molecular biology. *BioEssays* **19**, 519–525 (1997).
50. Schmitt, J., Dudley, S. A. & Pigliucci, M. Manipulative approaches to testing adaptive plasticity: phytochrome-mediated shade-avoidance responses in plants. *Am. Nat.* **154**, S43–S54 (1999).
51. Ballare, C. L. & Scopel, A. L. Phytochrome signalling in plant canopies: testing its population-level implications with photoreceptor mutants of *Arabidopsis*. *Funct. Ecol.* **11**, 441–450 (1997).
52. Pigliucci, M. & Schmitt, J. Genes affecting phenotypic plasticity in *Arabidopsis*: pleiotropic effects and reproductive fitness of photomorphogenic mutants. *J. Evol. Biol.* **12**, 551–562 (1999).
53. Schmitt, J., McCormac, A. C., Smith, H. A test of the adaptive plasticity hypothesis using transgenic and mutant plants disabled in phytochrome-mediated elongation responses to neighbors. *Am. Nat.* **146**, 937–953 (1995).
54. Smith, H. Signal perception, differential expression within multigene families and the molecular basis of phenotypic plasticity. *Plant Cell Env.* **13**, 585–594 (1990).
55. McSteen, P. & Hake, S. Genetic control of plant development. *Curr. Op. Biotechnol.* **9**, 189–195 (1998).
56. Puruggannan, M. D. The molecular genetics of regulatory genes. *Mol. Ecol.* **9**, 1451–1462 (2000).
57. Meyerowitz, E. M. Plants, animals and the logic of development. *Trends Biochem. Sci.* **24**, M65–M68 (1999).
58. Ting, C. T., Tsaui, S. C., Wu, M. L. & Wu, C. I. A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science* **282**, 1501–1504 (1998).
59. Puruggannan, M. D., Rounsley, S. D., Schmidt, R. J. & Yanofsky, M. F. Molecular evolution of flower development: Diversification of the plant MADS-box regulatory gene family. *Genetics* **140**, 354–356 (1995).
60. Mitchell-Olds, T. The molecular-basis of quantitative genetic-variation in natural-populations. *Trends Ecol. Evol.* **10**, 324–328 (1995).
61. Alonso-Blanco, C. & Koornneef, M. Naturally occurring variation in *Arabidopsis*: an underexploited resource for plant genetics. *Trends Plant Sci.* **5**, 22–29 (2000).
62. Alonso-Blanco, C., Blankestijn-de Vries, H., Hanhart, C. J. & Koornneef, M. Natural allelic variation at seed size loci in relation to other life history traits of *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **96**, 4710–4717 (1999).
63. Alonso-Blanco, C., El-Assal, S. E., Coupland, G. & Koornneef, M. Analysis of natural allelic variation at flowering time loci in the Landsberg *erecta* and Cape Verde Isles ecotypes of *Arabidopsis thaliana*. *Genetics* **149**, 749–764 (1998).
64. Swarup, K. *et al.* Natural allelic variation identifies new genes in the *Arabidopsis* circadian system. *Plant J.* **20**, 67–77 (1999).
65. Robson, P. R. H., McCormac, A. C., Irvine, A. S. & Smith, H. Genetic engineering of harvest index in tobacco through overexpression of a phytochrome gene. *Nature Biotechnol.* **14**, 995–998 (1996).
66. Olsen, J. E. *et al.* Ectopic expression of oat phytochrome A in hybrid aspen changes critical daylength for growth and prevents cold acclimatization. *Plant J.* **12**, 1339–1350 (1997).
67. Donoghue, M. J. & Mathews, S. Duplicate genes and the root of angiosperms, with an example using phytochrome sequences. *Molecular Phylogenet. Evol.* **9**, 489–500 (1998).
68. Mathews, S. & Donoghue, M. J. The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science* **286**, 947–950 (1999).
69. Quail, P. H. An emerging molecular map of the phytochrome. *Plant Cell Environment* **20**, 657–665 (1997).
70. Whitelam, G. C. & Devlin, P. F. Roles of different phytochromes in *Arabidopsis* photomorphogenesis. *Plant Cell Env.* **20**, 752–758 (1997).
71. Robson, P. R. H. & Smith, H. Fundamental and biotechnological applications of the phytochromes. *Plant Cell Env.* **20**, 831–839 (1997).

Acknowledgements

I thank the following for information and comment: J. Chory, M. Furuya, J. C. Lagarias, P. H. Quail, J. Schmitt, P.-S. Song. The author's research on shade avoidance was supported by the UK Natural Environment Research Council, the UK Biotechnology and Biological Sciences Research Council and the Council of the European Commission. The literature survey for this article was completed on 31 March 2000.

Correspondence should be addressed to H.S. (e-mail: has@le.ac.uk).



Id2 is a retinoblastoma protein target and mediates signalling by Myc oncoproteins

Anna Lasorella^{*†}, Michela Nosedà^{*}, Mercedes Beyna^{*} & Antonio Iavarone^{*†‡}

^{*} Department of Neurology, [‡] Department of Developmental and Molecular Biology, and [†] Comprehensive Cancer Center, Albert Einstein College of Medicine, Bronx, New York 10461, USA

In mammalian cells, Id proteins coordinate proliferation and differentiation. Id2 is a dominant-negative antagonist of basic helix–loop–helix transcription factors and proteins of the retinoblastoma (Rb) family. Here we show that *Id2–Rb* double knockout embryos survive to term with minimal or no defects in neurogenesis and haematopoiesis, but they die at birth from severe reduction of muscle tissue. In neuroblastoma, an embryonal tumour derived from the neural crest, Id2 is overexpressed in cells carrying extra copies of the *N-myc* gene. In these cells, Id2 is in molar excess of the active form of Rb. The overexpression of Id2 results from transcriptional activation by oncoproteins of the Myc family. Cell-cycle progression induced by Myc oncoproteins requires inactivation of Rb by Id2. Thus, a dual connection links Id2 and Rb: during normal cell-cycle, Rb prohibits the action of Id2 on its natural targets, but oncogenic activation of the Myc–Id2 transcriptional pathway overrides the tumour-suppressor function of Rb.

Ids regulate differentiation¹ through sequestration of basic helix–loop–helix (bHLH) transcription factors, and consequent inhibition of their ability to bind DNA². Although all Id proteins are viewed as positive regulators of cell-cycle progression, this role has been firmly established only for one member of the Id family, Id2 (refs 1, 3, 4). Only Id2, and not the other members of the family, Id1 and Id3, is able to disrupt the antiproliferative effects of tumour suppressor proteins of the Rb family (the ‘pocket’ proteins: Rb, p107 and p130), thus allowing cell-cycle progression. This function correlates with the ability of Id2, but not Id1 and Id3, to associate physically with active, hypophosphorylated forms of the pocket proteins *in vitro* and *in vivo*. By inactivating Rb, Id2 is also able to abolish the function of another growth-inhibitory protein, p16, that operates upstream to Rb^{3,4}.

The *Rb*-null phenotype is lethal by embryonic day (E) 14.5 (refs 5–7) because of widespread proliferation, defective differentiation and apoptosis in the nervous system and haematopoietic precursors^{8,9}. As Id2 is expressed in these cell types at the time that *Rb*-null embryos die^{10–12}, we hypothesize that, if Id2 is a natural target of Rb, manifestation of the *Rb*-mutant phenotype might require intact *Id2*.

Disruption of the Rb pathway (which also includes cyclin D, cdk4/6 and p16) is a hallmark of cancer and it is widely accepted that normal Rb function must be removed, one way or another, in all human tumours^{13,14}. Therefore, we set out to determine whether tumour cells deregulate Id2 to bypass the Rb pathway. Correct expression of Id2 is essential to regulate proliferation and differentiation of the neural crest, thus neural crest precursor cells might be sensitive to inappropriate expression of Id2 (ref. 15). In humans, neoplastic transformation of neural crest precursors during embryogenesis causes neuroblastoma¹⁶. Interestingly, genetic alterations of *Rb*, *cyclin D*, *cdk4/6* or *p16* are absent in neuroblastoma^{17–23}. The genetic hallmark of neuroblastoma is amplification of the gene for a member of the Myc family of proto-oncogenes, *N-myc*. Resembling enforced expression of Id2, Myc overexpression is sufficient to bypass the Rb–p16 growth-inhibitory pathway, in spite of persistent hypophosphorylated Rb^{24,25}. Consequently, Myc activation may release the pressure to mutate components of the Rb–p16 pathway during tumorigenesis^{24,26}.

In this work, we have intercrossed *Rb* and *Id2* mutant mice and we have shown a genetic interaction between *Rb* and *Id2* during development. In addition, through our analysis of the involvement of Id2 in neuroblastoma, we have discovered an unexpected link between Myc oncoproteins and Id2. Activation of *N-myc* proto-oncogene in neuroblastoma causes an increase in Id2 to more than the otherwise active, hypophosphorylated Rb in these cells. We find that this effect is the result of inappropriate activity of a transcriptional network where Myc transcription factors increase expression of Id2 to bypass the Rb block and drive progression of the cell cycle.

Mutation of *Id2* rescues *Rb* mutant embryos

To test whether elimination of Id2 would affect survival and/or phenotypic defects of *Rb*^{−/−} embryos, we intercrossed *Rb*^{+/−} animals carrying variable *Id2* backgrounds (*Id2*^{+/+}, *Id2*^{+/-}, *Id2*^{−/−}; Table 1). Intercrosses from *Id2*^{+/+} *Rb*^{+/−} parents generated no *Rb*^{−/−} pups. At E15.5, *Rb*^{−/−} embryos were found in advanced state of resorption confirming that death occurred by E13.5–E14.5 (refs 5–7). However, intercrosses between *Id2*^{+/-} *Rb*^{+/−} and *Id2*^{−/−} *Rb*^{+/−} parents generated 100% of expected *Id2*^{−/−} *Rb*^{−/−} pups and 50% of expected *Id2*^{+/-} *Rb*^{−/−} pups. These results indicate that loss of *Id2* rescued the embryonic lethality of homozygous *Rb* mutants in a dose-dependent manner.

The *Id2*^{−/−} *Rb*^{−/−} and *Id2*^{+/-} *Rb*^{−/−} mice were stillborn and phenotypically identifiable at birth because of a characteristic hunchback phenotype (Fig. 1a, b). Alizarin red/alcian blue staining of skeletons revealed an abnormal curvature of the vertebral column and a deformed rib cage (Fig. 1c, d). Gross morphological analysis showed dramatic reduction of skeletal muscle tissue. Histological examination and immunostaining for myosin heavy chain (MHC) in

Table 1 Frequency of *Rb*^{−/−} offspring at birth

Cross	Total neonates	<i>Rb</i> ^{−/−} observed (<i>Rb</i> ^{−/−} expected)		
		<i>Id2</i> ^{+/+}	<i>Id2</i> ^{+/-}	<i>Id2</i> ^{−/−}
I. <i>Id2</i> ^{+/+} <i>Rb</i> ^{+/−} × <i>Id2</i> ^{+/+} <i>Rb</i> ^{+/−}	81	10 (10)	5 (10)	—
II. <i>Id2</i> ^{+/+} <i>Rb</i> ^{+/−} × <i>Id2</i> ^{+/-} <i>Rb</i> ^{+/−}	13	3 (3)	—	—
III. <i>Id2</i> ^{+/+} <i>Rb</i> ^{+/−} × <i>Id2</i> ^{−/−} <i>Rb</i> ^{+/−}	113	—	—	0 (28)

Numbers represent neonates derived from 13 litters (cross I), 2 litters (cross II) and 19 litters (cross III).

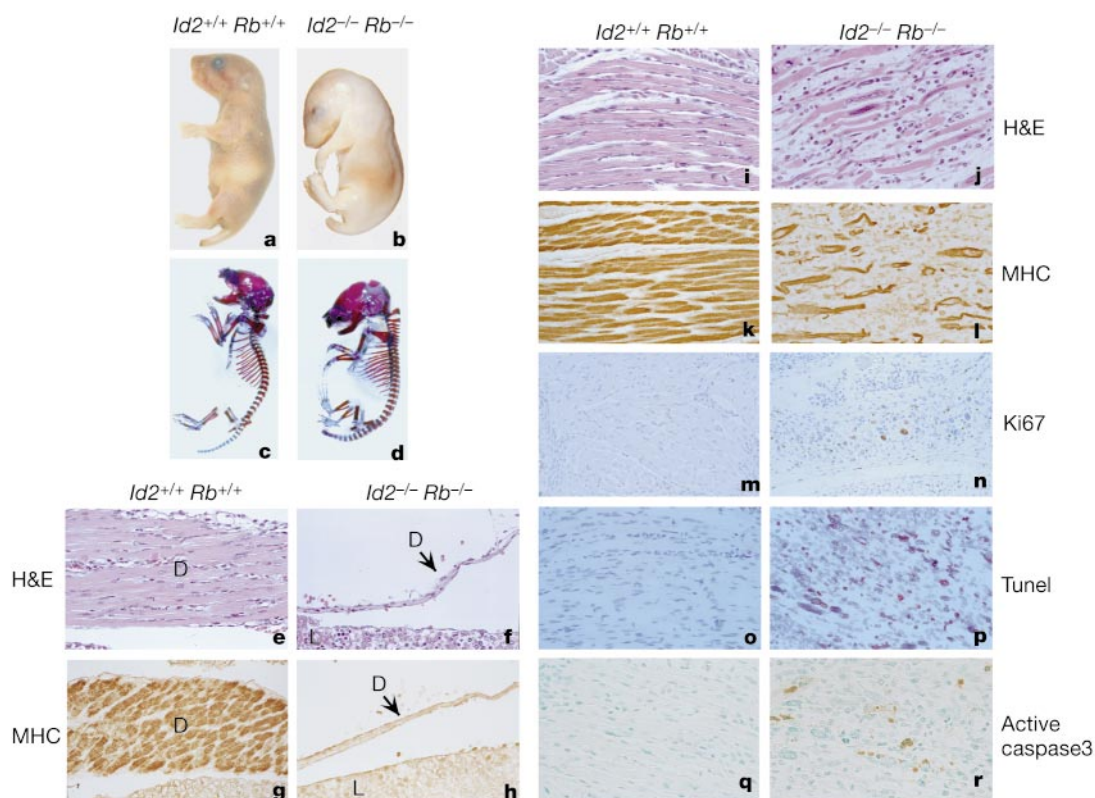


Figure 1 Abnormal myogenesis in *Id2*^{-/-} *Rb*^{-/-} neonates. **a–d**, Double mutant mice display altered posture at birth (hunchback). Diaphragms (**e–h**) and axial muscles (**i–l**) were stained with haematoxylin and eosin (H & E; **e, f, i, j**), immunostained with myosin heavy chain (MHC; **g, h, k, l**) and Ki67 (**m, n**), analysed by TUNEL assay (**o, p**) and

immunostained with active caspase-3 (**q, r**). Sections shown in **m–p** and **q–r** were counterstained with haematoxylin and methyl green, respectively. Arrow, the abnormal diaphragm of *Id2*^{-/-} *Rb*^{-/-} mice; D, diaphragm; L, liver. Original magnification: $\times 40$ (**e–l, o–r**); $\times 25$ (**m, n**).

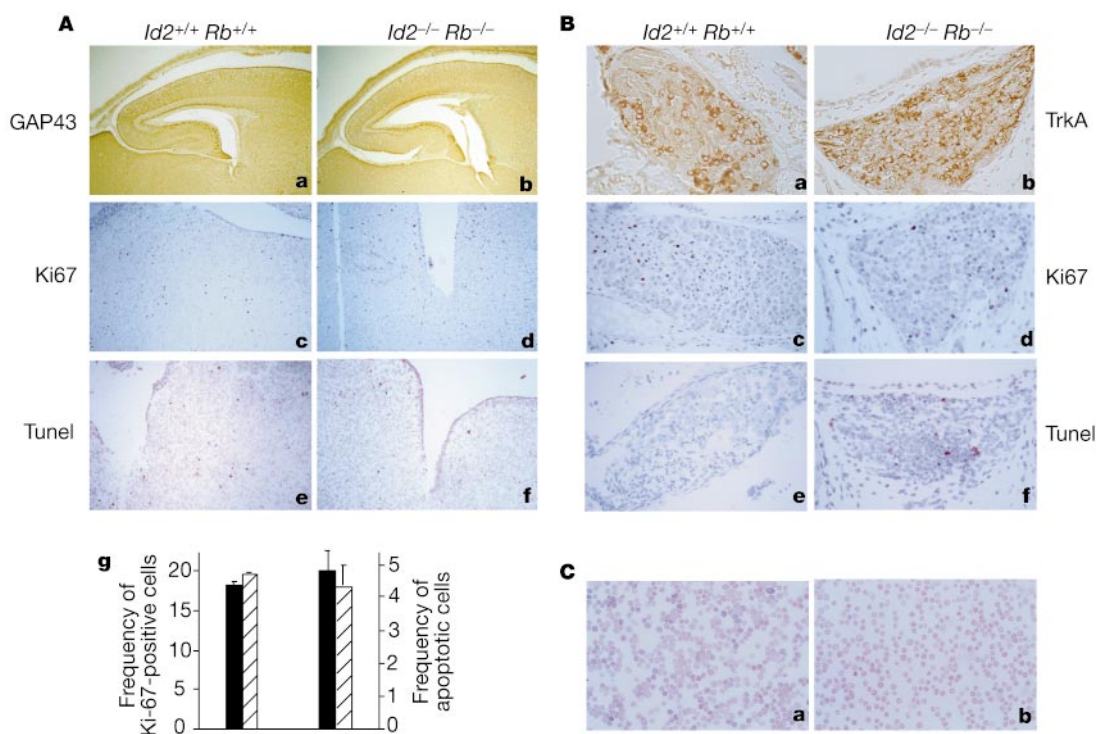


Figure 2 Normal neurogenesis and erythropoiesis in *Id2*^{-/-} *Rb*^{-/-} neonates. Whole brain sections were immunostained with GAP43 (**Aa, b**). Dorsal root ganglia were immunostained with TrkA (**Ba, b**). Forth ventricle of the brain (**Ac–f**) and dorsal root ganglia (**Bc–f**) were immunostained with Ki67 (**Ac, d, Bc, d**) and analysed by TUNEL assay

(**Ae, f, Be, f**). **Ac–f** and **Bc–f** were counterstained with haematoxylin. Original magnification: $\times 2.5$ (**Aa, b**); $\times 10$ (**Ac–f**); $\times 25$ (**Ba–f**). **Ag**, Frequency of proliferating and apoptotic cells in the CNS. Black bars, wild type ($n = 3$), hatched bars, *Id2*^{-/-} *Rb*^{-/-} ($n = 3$). **Ca, b**, Peripheral blood smears.

regions normally occupied by respiratory muscles showed complete absence of muscle fibres (Fig. 1e–h). Therefore, loss of respiratory muscle function at birth is the most likely cause of death of *Id2*–*Rb* double mutant mice. Histological analysis and MHC immunostaining of axial and limb regions of *Id2*^{−/−} *Rb*^{−/−} pups showed that myotubes were widely dispersed compared with the compact arrangement of muscle cells from control mice. However, expression of MHC in the residual myotubes was relatively normal (Fig. 1i–l). In the absence of *Id2* and *Rb*, myotubes were frequently positive for the proliferating antigen Ki67 (Fig. 1m, n) and showed an increased rate of apoptosis, as demonstrated by TUNEL assay and immunostaining for the active form of caspase-3 (Fig. 1o–r).

The phenotypic hallmarks responsible for lethality of *Rb*^{−/−} mice during embryogenesis, neurological and erythroid abnormalities, were virtually absent in *Id2*^{−/−} *Rb*^{−/−} neonates (Fig. 2). In the central nervous system (CNS), immunostaining for three post-mitotic neuronal markers (GAP43, MAP2, neurofilament 160) indicated that normal neuronal differentiation occurred in *Id2*^{−/−} *Rb*^{−/−} mice throughout development (Fig. 2A, a and b; and data not shown). *Id2*^{−/−} *Rb*^{−/−} brains showed no evidence of inappropriate proliferating activity (Fig. 2A, c and d) or of increased apoptosis (Fig. 2A, e and f), as determined by quantitative immunodetection of Ki67 and TUNEL assay, respectively (Fig. 2A, g). In the peripheral nervous system we analysed dorsal root ganglia (DRG) and trigeminal ganglia (Fig. 2B; and see Supplementary Information), two tissues that are profoundly degenerated in *Rb* mutant embryos^{5–7,9}. Double-mutant DRG showed a near-normal expression of TrkA, a late differentiation marker of peripheral neurons that is lost in *Rb* mutant embryos^{9,27} (Fig. 2B, a and b). Similarly, Ki67 immunostaining (Fig. 2B, c and d) was indistinguishable from normal controls. However, we saw a moderate increase of the number of apoptotic cells in *Id2*^{−/−} *Rb*^{−/−} DRG (Fig. 2B, e and f). Peripheral blood smears from *Id2*^{−/−} *Rb*^{−/−} mice were generally indistinguishable from their normal littermates (Fig. 2C, a and b). Occasionally, we found less than 1% nucleated red blood cells.

Regulation of Id2–pocket protein complexes

We asked whether endogenous Id2 and pocket proteins interact in normal cells and whether these interactions are regulated during cell cycle. Id2 immunoprecipitates from human diploid fibroblasts (TIG-3) were analysed for pocket proteins by western blot (Fig. 3). We prepared TIG-3 extracts from serum-starved cells (0 h; cell-cycle phase G0) and from cells stimulated by serum for 4 h (G1) and 14 h (entering S phase) (data not shown). Id2 was low but detectable in

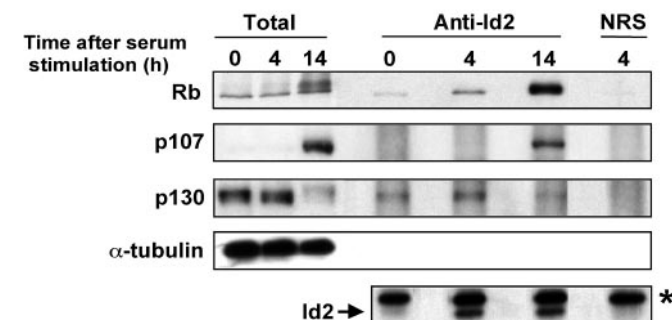


Figure 3 Cell-cycle analysis of complexes of Id2 with pocket proteins. TIG-3 fibroblasts were synchronized in G0 by serum starvation and stimulated to re-enter the cell cycle by addition of 20% serum. Lysates were prepared from cells harvested at 0 h, 4 h and 14 h after serum addition, and proteins were immunoprecipitated with polyclonal anti-Id2 antibody (α-Id2) or normal rabbit serum (NRS). Immune complexes were analysed by western immunoblot for Rb, p107, p130, α-tubulin and Id2. The same antibodies were used to detect proteins from total cellular lysates (Total). Arrow indicates migration of immunoprecipitated Id2, just below a nonspecific band (asterisk).

quiescent cells and was rapidly induced by serum, with persistence of high amounts until the beginning of S phase (Fig. 3). Id2 immunoprecipitates were free of abundant cellular proteins such as α-tubulin, and pocket proteins were never precipitated by normal rabbit serum (Fig. 3; and data not shown).

Id2 bound exclusively to the hypophosphorylated forms of pocket proteins. In quiescent cells, the low amounts of Id2 were mainly associated with p130, and complexes between Id2 and p107 were undetectable in G0 and during transit through G1. The increase of Id2 4 h after addition of serum resulted in only moderate accumulation of Id2–Rb and Id2–p130 complexes. However, Id2–Rb complexes increased significantly at the beginning of S phase. At this time, the Id2–p107 interaction was also detectable, whereas Id2–p130 complexes were virtually absent.

Id2 overrides Rb in neuroblastomas with *N-myc* amplification

Rb is required for normal development, and integrity of the Rb pathway is essential to prevent tumorigenesis. Hence, expression of Id2 was analysed in ten human neuroblastoma cell lines (Fig. 4a), six of which had extra copies of the *N-myc* gene (NGP, LAN1, LAN15n, IMR32, SMS-KCNR, SK-N-DZ). Five of these six had 20–30-fold more Id2 than the neuroblastoma cell lines without amplification of

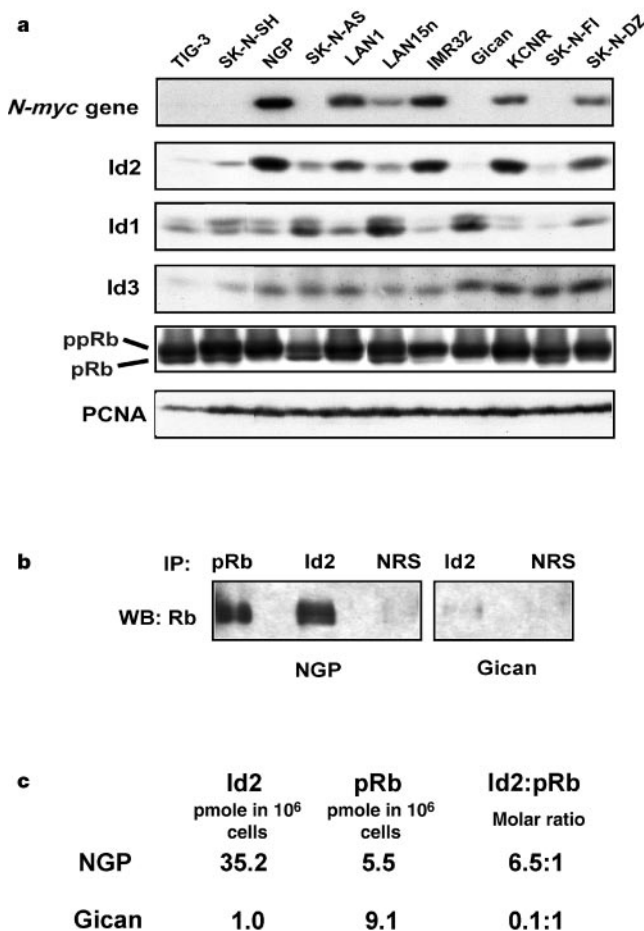


Figure 4 Id2 expression in neuroblastoma cell lines. **a**, Southern blot for *N-myc* and Western blot for Id2, Id1, Id3, Rb and PCNA from ten exponentially growing neuroblastoma cell lines and TIG-3 fibroblasts. pRb, hypophosphorylated Rb, ppRb, hyperphosphorylated Rb. **b**, Immunoprecipitations from cell lysates of NGP cells, which have amplified *N-myc*, and Gican cells, which do not. Immunoprecipitation (IP) used anti-Id2 antibodies or normal rabbit serum (NRS) and was followed by Western blot (WB) for Rb. One fifth of NGP lysates used in anti-Id2 and NRS IP was quantitatively immunoprecipitated with the anti-hypophosphorylated Rb antibody G99-549 (pRb). **c**, Molar amounts of Id2 and hypophosphorylated Rb in NGP and Gican.

N-myc (SK-N-SH, SK-N-AS, Gican, SK-N-FI). Expression of Id2 in neuroblastoma cell lines without *N-myc* amplification was comparable to TIG-3 fibroblasts (Fig. 4a). The only neuroblastoma cell line with *N-myc* amplification but lower expression of Id2 was LAN15n, a LAN1 derivative with fewer copies than parental cells of the *N-myc* gene and reduced expression of N-Myc messenger RNA. Expression of Id1 tended to be the opposite of Id2 with more in cell lines without amplification of *N-myc*. Id3 showed minimal differences among neuroblastoma cell lines (Fig. 4a). Southern blot analysis of p16 and cdk6, northern blot analysis of p16 and western blot analysis of p16, cyclin D1 and cdk4 indicated that, as reported by others^{17–23}, neuroblastoma cell lines carried no alterations of the known components of the Rb pathway (data not shown).

Although the degrees of Rb phosphorylation varied between neuroblastoma cells, each cell line contained detectable amounts of hypophosphorylated Rb (Fig. 4a). Next, we measured the fraction of hypophosphorylated Rb recoverable as a stable complex with Id2 in neuroblastoma cell lines with or without *N-myc* amplification (Fig. 4b). Id2 immunoprecipitates from neuroblastoma cells contained only hypophosphorylated Rb, as shown by co-migration with Rb precipitated by an antibody specific for hypophosphorylated Rb (Fig. 4b). In NGP, a cell line with *N-myc* amplification, a minimum of 20% of total hypophosphorylated Rb was found in a complex with Id2. However, in Gican (without *N-myc* amplification) only 0.5–1% hypophosphorylated Rb was bound to Id2. To bypass growth suppression by Rb, the molar amount of Id2 should exceed the amount of hypophosphorylated Rb in neuroblastoma cells with *N-myc* amplification. We measured the Id2:Rb molar ratio in NGP and Gican (Fig. 4c). Id2 was tenfold less than hypophosphorylated Rb in Gican. The Id2:Rb ratio was reversed in NGP where the molar amount of Id2 exceeded hypophosphorylated Rb by a factor of 6.5. This ratio is within the range of ectopic Id2:Rb

stoichiometry in Saos-2 cells where expression of Id2 overcomes cell-cycle arrest by Rb³.

Id2 is an N-Myc and c-Myc target

To test whether overexpression of Id2 in neuroblastoma is the result of direct control of *Id2* gene expression by N-Myc, *N-myc* complementary DNA was fused with oestrogen receptor (ER) cDNA and the resulting N-Myc-ER fusion protein was expressed in NIH 3T3 cells (Fig. 5a). After starving these cells of serum, activation of N-Myc-ER by 4-hydroxytamoxifen (4-OHT) caused entry into S phase (Fig. 5b). This treatment also caused a fast, strong and sustained increase of *Id2* mRNA (Fig. 5c). *Id2* was not induced by 4-OHT treatment of parental NIH 3T3 or NIH 3T3 that had been infected with an empty virus (data not shown). Induction of *Id2* mRNA by N-Myc-ER activation was maintained in the presence of cycloheximide (see Supplementary Information), and resulted into rapid accumulation of Id2 protein with no detectable change of abundance or phosphorylation of Rb (Fig. 5d).

We sought to determine whether targeting Id2 expression is a common attribute of Myc family members. First, we asked whether expression of c-Myc and Id2 were correlated during the proliferative response of normal human fibroblasts to serum. c-Myc was rapidly induced by serum (15 min) and immediately followed (30 min) by upregulation of Id2. Between 8 and 16 h, both c-Myc and Id2 were downregulated (Fig. 6a). Repression of c-Myc is a classical gene response to the anti-mitogenic cytokine TGF- β (refs 28, 29). Accordingly, TGF- β also repressed Id2 in a variety of cell types (wild-type and *p15*^{-/-} MEFs, Mv1Lu mink lung epithelial cells, and HaCaT human keratinocytes; Fig. 6b, c and data not shown). Interestingly, c-Myc and Id2 were the only cell-cycle-related genes modified by TGF- β in wild-type and *p15*^{-/-} MEFs (Fig. 6b and data not shown for G1 cyclins, CDKs and CDK inhibitors).

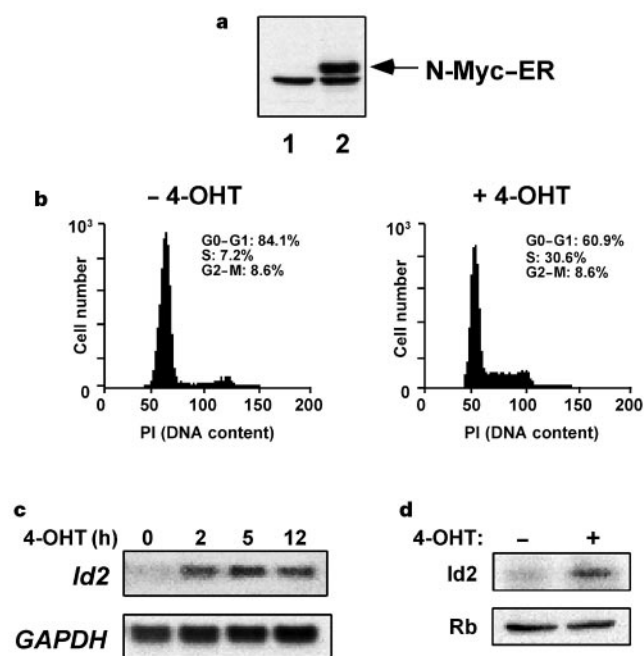


Figure 5 *Id2* is an N-Myc target gene. **a**, N-Myc western blot with lysates from NIH 3T3 infected with pBabe-puro (lane 1) or pBabe-puroN-mycER retroviruses (lane 2). **b**, NIH cells containing *N-myc*-ER were starved of serum (- 4-OHT), then treated with 4-hydroxytamoxifen (4-OHT) for 12 h (+ 4-OHT). Cell-cycle progression was analysed by flow cytometry. **c**, After 4-OHT treatment for the indicated times, total RNA was analysed by northern blot for *Id2* and *GAPDH*. **d**, Lysates from *N-myc*-ER containing cells, untreated (-) and treated with 4-OHT (+), were analysed by western blot for Id2 and Rb.

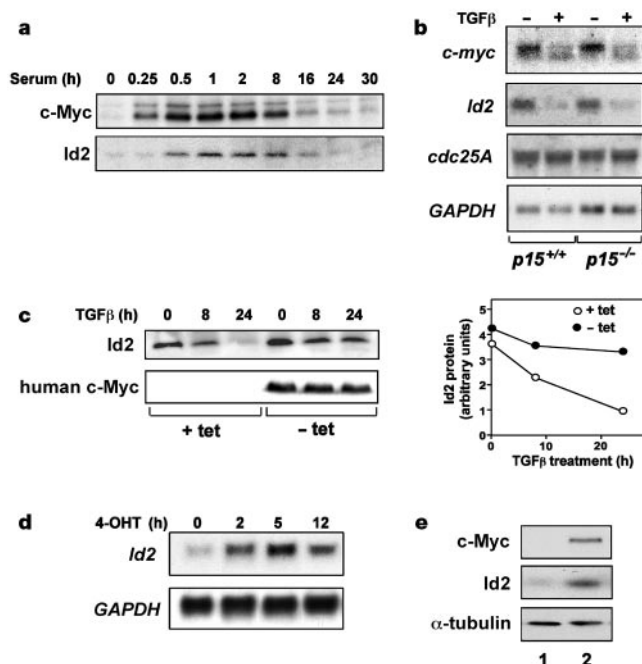


Figure 6 *Id2* is a c-Myc target gene. **a**, Western blot from serum-starved TIG-3 that were re-stimulated by serum for the indicated times. **b**, Northern blot of wild-type and *p15*^{-/-} MEFs left untreated (-) and treated with TGF β (+) for 8 h. **c**, Western blot of Mv1Lu expressing inducible c-Myc treated with (+ tet) or without (- tet) tetracycline followed by treatment with TGF β for the indicated times. Densitometric quantification of Id2 is shown on the right. **d**, Northern blot of total RNA from serum-starved Rat-1 cells containing *c-Myc*-ER treated with 4-OHT for the indicated times. **e**, c-Myc and Id2 proteins in MEFs infected with LZRS-GFP (lane 1) or LZRS-c-Myc-GFP (lane 2) retroviruses.

Downregulation of *Id2* by TGF- β was essentially abolished in Mv1Lu expressing a tetracycline-regulatable *c-Myc* gene, indicating that loss of *c-Myc* is indispensable for efficient repression of *Id2* (Fig. 6c). Furthermore, using a *c-Myc*-ER fusion in Rat-1 cells (Fig. 6d), MEF cells and NIH 3T3 cells (see Supplementary Information), *c-Myc* activation by 4-OHT resulted in efficient induction of *Id2* expression. Likewise, expression of wild-type *c-Myc* in MEFs elevated *Id2* (Fig. 6e).

Activation of the *Id2* promoter by Myc oncoproteins

To ascertain whether Myc oncoproteins upregulate *Id2* at the transcriptional level, we cloned the human *Id2* promoter. The 5' region of human *Id2* contains a cluster of three E-boxes that is fully conserved in mouse *Id2* (Fig. 7a)³⁰. The three E-boxes are scattered throughout a region of 400 base pairs (bp) that has 80% of nucleotides identical to the mouse gene (data not shown). Two of the three E-boxes are high-affinity Myc-binding sites (CACATG at position -1,880 and CACGTG at position -1,590; Fig. 7a). To determine whether the E-box cluster in the *Id2* promoter binds *c-Myc* *in vivo* during the proliferation response of human fibroblasts to serum, we used chromatin crosslinking³¹. *c-Myc* did not

bind to *Id2* promoter in quiescent fibroblasts but it did after stimulation by serum (Fig. 7b). Furthermore, *c-Myc* and N-Myc transactivated a 2,900 bp *Id2* promoter fragment in a dose-dependent manner (Fig. 7c, d). Induction of the *Id2* reporter by Myc transcription factors required the E-box cluster, as shown by loss of Myc responsiveness of *Id2* promoter constructs lacking this cluster (Fig. 7c, d for pGId2-1330 and data not shown for additional 5' deletion constructs). To test whether the E-boxes in the *Id2* promoter could mediate transcriptional activation by Myc oncoproteins, an *EcoRI* fragment containing wild-type or mutant Myc-binding sites was placed in front of a minimal *Id2* promoter (pGId2-*EcoRI* and pGId2-*EcoRI*m, Fig. 7a). The reporter plasmid pGId2-*EcoRI* underwent dose-dependent activation by *c-Myc* and N-Myc but mutations of the Myc-binding sites in pGId2-*EcoRI*m completely abolished this activation (Fig. 7c, d).

Myc requires *Id2* to bypass Rb

Wild-type and *Id2*-null MEFs were infected with *c-Myc*-ER and N-Myc-ER retroviruses and starved of serum. Activation of *c-Myc*-ER and N-Myc-ER by 4-OHT treatment induced entry into S phase of wild-type MEFs. This effect was absent in *Id2*-null MEFs,

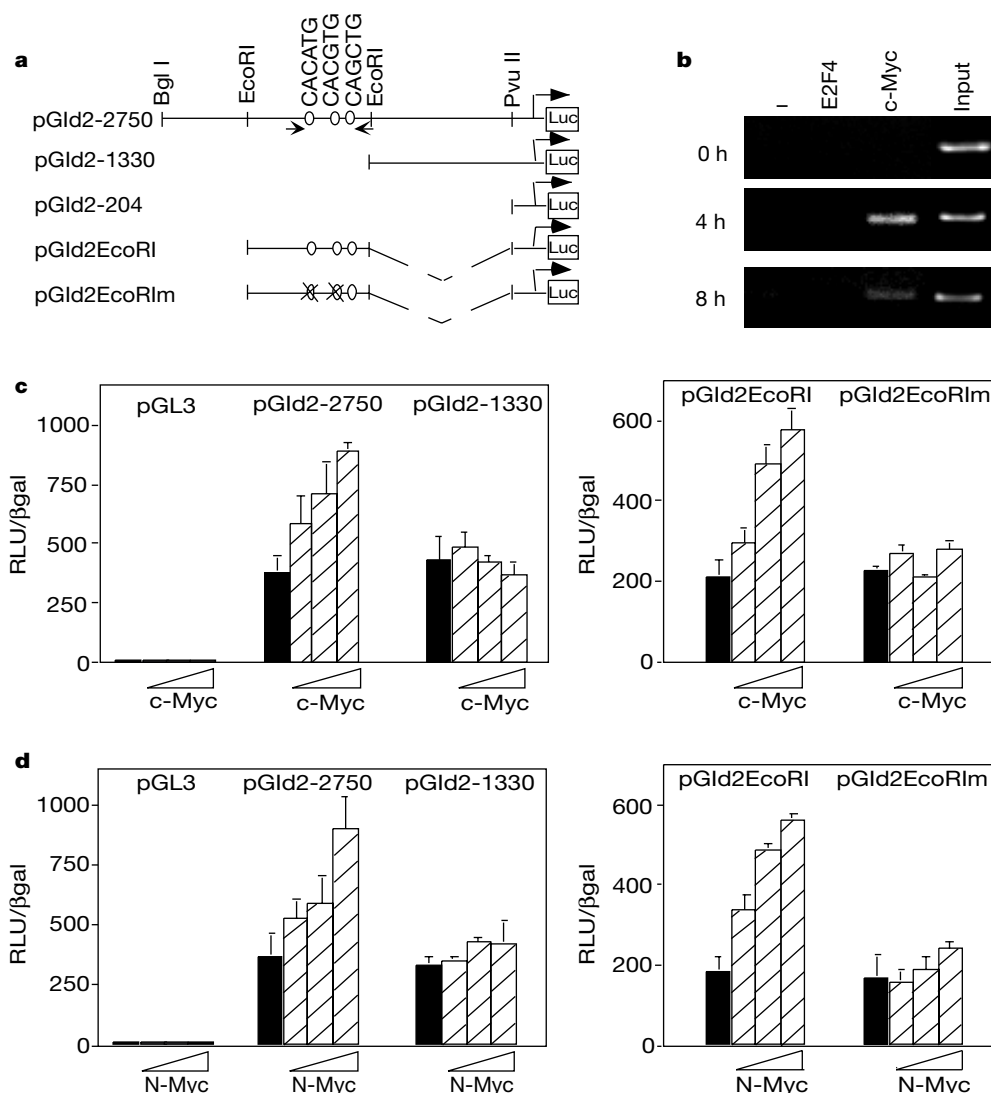


Figure 7 Binding and transactivation of *Id2* promoter by Myc transcription factors. **a**, *Id2* promoter-luciferase constructs. Arrows indicate primers of PCR reactions in chromatin immunoprecipitation assay. **b**, Crosslinked chromatin was prepared from serum-starved TIG-3 (0 h) and from cells re-stimulated by serum for 4 h and 8 h. Chromatin was incubated without antibody (-), with anti-E2F4 (E2F4) or with anti-c-Myc (c-Myc)

antibodies. Immunoprecipitates were analysed by PCR. Input is 0.02% of total input chromatin. **c**, **d**, 293 T cells were transfected with *Id2*-luciferase constructs in the absence (black bars) or presence (hatched bars) of increasing amounts of pcDNA3-c-Myc (**c**) and pcDNA3-N-Myc (**d**). Luciferase activity is the relative luminescence units normalized to β -galactosidase activity.

but was restored when Id2 was re-expressed before infection with Myc-ER retroviruses (Fig. 8a, b). Expression of wild-type Myc prevented cell-cycle exit upon serum withdrawal in *Id2*^{+/+} MEFs but was ineffective in *Id2*-null MEFs (Fig. 8c). Similarly, ability of Myc to promote colony formation at low cell density was abolished in the absence of Id2 (data not shown). Interestingly, cell-cycle progression by Myc-ER activation and wild-type Myc expression was restored in MEFs from *Id2*^{-/-} *Rb*^{-/-} embryos, indicating that Myc requires Id2 to bypass Rb (Fig. 8a-c). Another Myc function, the cooperation with Ras for cellular transformation, was neutralized by loss of Id2, yet *Id2*^{-/-} MEFs could still be transformed by the combination of *E1A* and *ras* oncogenes (Fig. 8d). However, Myc-induced apoptosis was not compromised in *Id2*-null MEFs (Fig. 8e).

Discussion

We show that Id2 function is essential for lethality and for the neurogenic and haematopoietic abnormalities caused by loss of Rb in the developing embryo. To our knowledge, this is the first time that a protein partner of Rb, distinct from E2F transcription factors, is shown to be essential for Rb function *in vivo*. Although our results clearly identify Id2 as a key target of Rb in normal cells, they also indicate that this scheme is subverted in neural cancer. There, the large amounts of Id2 that are generated as a direct consequence of genetic amplification of the *N-myc* oncogene neutralize the otherwise intact Rb pathway in these cells. Thus, Id2, a protein required to

maintain the timing of differentiation in mammalian development, is transformed into the effector of a powerful oncogene in human cancer.

The hallmarks of the *Rb* mutant phenotype, ectopic cellular proliferation, loss of differentiation and cell death, correspond to the consequences of uncontrolled Id2 activity^{3,4,15,32,33}. Our analysis of *Rb*-*Id2* double mutant mice highlights Id2 as a downstream target of Rb during neurogenesis and erythropoiesis. Furthermore, the identification in normal cells of cell-cycle-regulated associations between Id2 and pocket proteins indicates that suppression of Id2 function by Rb is likely to be a general mechanism by which pocket proteins control progression of the cell cycle. The observation that Id2 (but not other Id proteins) is expressed in differentiating neuronal and haematopoietic cells^{11,34}, and that these cell types are able to re-enter inappropriate proliferation and ultimately apoptosis in the absence of Rb is consistent with the idea that maintenance of the post-mitotic state requires physiological control of Id2 by Rb. Therefore, Id2 must target other factors during development and normal cell-cycle progression and these targets must be activated following Rb-mediated inactivation of Id2.

So far, rescue of neurogenic and erythropoietic defects causing embryonic lethality of *Rb* mutant embryos has been accomplished by introduction of an *Rb*ox minigene⁸, by mutation of *Id2* (this work) and, partially, by mutation of *E2F-1* (ref. 35). Clearly, in all three settings Rb is indispensable in muscle differentiation and survival. This indicates that the role of Rb in muscle differentiation may be profoundly different from its function in neurogenesis and erythropoiesis. In addition, evidence that Id2 is neither expressed in differentiating skeletal muscle during embryogenesis nor in myotubes from adult mice (ref. 36; and see Supplementary Information) indicates that physiological Rb targets in muscle cells have to be identified.

The large increase of Id2 in neuroblastoma cell lines (and primary tumours; A.L. and A.I., unpublished data) provides tumour cells with an epigenetic mechanism to bypass the tumour-suppressor function of Rb, in the absence of genetic alterations of the Rb pathway. We speculate that other unknown mechanisms aimed at overriding the Rb pathway might operate in neuroblastoma cell lines without *N-myc* amplification/Id2 overexpression.

Our results provide compelling evidence in favour of Id2 as a physiologically relevant, direct target of Myc transcription factors. Two phenotypic hallmarks of Myc activation, the ability to promote cell-cycle entry in the absence of growth factors and the ability to cooperate with Ras to transform normal fibroblasts, are strictly dependent on the presence of Id2. Consequently, human neuroblastoma tumorigenesis, when sustained by *N-myc* gene amplification, is invariably associated with overexpression of Id2. Our experiments in which Myc-expressing retroviruses are introduced in genetically modified MEFs formally demonstrate a trivalent connection between Myc, Id2 and Rb where Myc must induce Id2 to overcome an Rb-imposed cell-cycle block. However, in the absence of Rb, Id2 also becomes dispensable indicating that, if the entire Id2-Rb pathway is removed, other Myc targets are sufficient to promote cell-cycle progression. The existence of additional Myc targets, possibly operating in a redundant manner with Id2, is also suggested by the apparent lack of major defects in *Id2*-null mice during embryogenesis³⁷ and by the intact ability of Myc to induce apoptosis in *Id2*-null MEFs (Fig. 8e).

We propose that small molecule inhibitors of Id2 in tumours carrying abnormal activation of the Myc-Id2 oncogenic network, such as neuroblastoma, might be useful to restore proliferation control by the Rb tumour-suppressor pathway. □

Methods

Mouse strains

Id2^{+/+} mice on 129/Sv genetic background³⁷ were crossed with *Rb*^{+/+} mice on C57BL/6 × 129/Sv mixed background⁶ to generate *Id2*^{+/+} *Rb*^{+/+}. These mice were subsequently

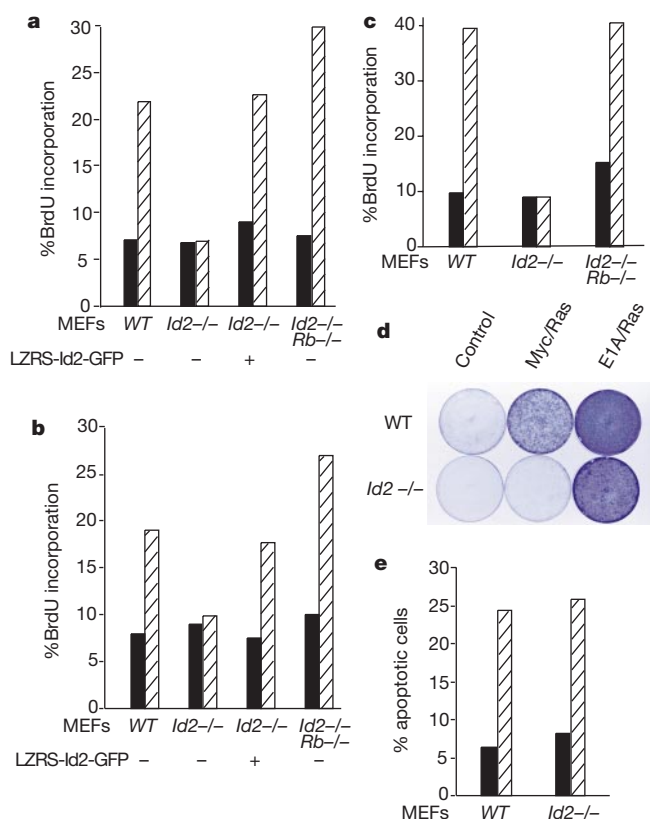


Figure 8 Functional analysis of the Id2-Rb pathway in Myc-induced proliferation, transformation and apoptosis. MEFs of the indicated genotypes were infected with pBabe-puroc-MycER (**a**), pBabe-puroN-mycER (**b**) and starved of serum. Where indicated, *Id2*^{-/-} MEFs were first infected with LZRS-Id2-GFP and subsequently with pBabe-puroMycER retroviruses. Myc-ER fusion proteins were activated by 4-OHT and cells were assayed for BrdU incorporation. Black bars, untreated; hatched bars, 4-OHT-treated. **c**, **e**, Cells infected with LZRS-GFP (black bars) or LZRS-c-Myc-GFP (hatched bars) were serum-starved and assayed for BrdU incorporation (**c**) and apoptosis (**e**). **d**, Transformation assay. MEFs were infected with retroviruses encoding the indicated proteins and transformed foci were stained with Giemsa after 12 days. These data are typical of several independent experiments.

intercrossed to generate founders for matings described in Table 1. Mice were genotyped routinely by polymerase chain reaction (PCR) of tail DNA as described^{6,37}.

Skeletal and histological analysis, immunohistochemistry and blood analysis

Alcian blue and alizarin red staining for cartilage and bones was performed as described³⁸. For histology, neonates were fixed in 10% formalin and embedded in paraffin, from which 4-µm serial sagittal sections were cut. Sections were stained with haematoxylin and eosin or used for immunohistochemistry with the indicated antibodies. Apoptosis was assayed using a commercial kit for TUNEL assay (Roche) and by active caspase-3 immunostaining (Pharmingen). Proliferation and apoptosis in the CNS was measured in multiple sections as a fraction of Ki67 and TUNEL-positive cells per unit area of tissue, respectively. Peripheral blood was drawn from tail vessels and smears were stained with Wright-Giemsa (Sigma).

Immunoprecipitations, immunoblot analysis and Id2/Rb quantifications

To analyse Id2 immunoprecipitates, cells were lysed as described²⁸ and complexes of Id2 with pocket proteins were immunoprecipitated with polyclonal antibodies against Id2 (C-20, Santa Cruz). Hypophosphorylated Rb was precipitated with monoclonal antibody G99-549 (ref. 3; Pharmingen). To measure the endogenous amounts of Id2 and hypophosphorylated Rb expressed by neuroblastoma cells, recombinant GST-Id2 and GST-Rb were produced in bacteria and purified. The chemiluminescent signal of known amounts of recombinant proteins was plotted against the signal given by serial dilutions of neuroblastoma cell extracts on the same western blot against Id2 and hypophosphorylated Rb. A list of antibodies used in immunohistochemistry and western blotting is available on request.

Cell culture

Cells were maintained in DMEM supplemented with 10% fetal bovine serum (Sigma). MEFs of the indicated genotypes were prepared from E13.5 embryos as described³⁹ and used within the first three passages in culture. Flow cytometry analysis was done as described³.

To generate high-titre retroviral stocks, Phoenix producer cells were used as described⁴⁰. For infection, subconfluent MEFs or NIH 3T3 cultures were incubated at 37 °C with viral supernatants in the presence of 4 µg ml⁻¹ polybrene (Sigma) for 12 h. Three rounds of infection were routinely performed to accomplish a 100% rate of infection. At the end of the last infection, the viral supernatant was diluted 1:3 with complete medium and left on the cells for 24 h. To activate Myc-ER fusion proteins in NIH 3T3 and MEFs, cells were treated with 4-OHT (250 nM, Sigma) after 36 h of starvation in DMEM containing 0.05% serum.

Incorporation of 5-deoxybromouridine (BrdU) was quantitated by immunofluorescence analysis performed on MEFs labelled with 10 µM BrdU (Roche) for 4 h. Myc-induced apoptosis was scored by visualization of nuclear condensation and blebbing of Hoechst-stained cells and validated by TUNEL assay of representative samples. The percentage of BrdU incorporation and apoptosis was determined by counting more than 1,000 cells.

Id2 promoter analysis

A human genomic library was screened with an Id2-specific cDNA probe and two phage clones were isolated and sequenced. A 3.0-kilobase (kb) BglI-NheI fragment was subcloned in pGL3 Basic (Promega) to generate pGId2-2750. The Id2 reporter plasmid pGId2-1330 was generated by 5' deletion of the larger promoter fragment pGId2-2750. Plasmids pGId2-EcoRI and pGId2-EcoRI_m were generated by placing a 900-bp EcoRI fragment upstream of a 204-bp minimal Id2 promoter. All plasmids harbour a 35-bp region downstream of the start site. Site-specific mutagenesis was performed by a PCR-based protocol and transfections and luciferase assays were done as described⁴¹. Luciferase activity was normalized to the expression of pCMV-LacZ cotransfected as an internal standard. PCR primers used for chromatin immunoprecipitation assays were 5'-TCTGTCTCCACTGTGGCAGCTAT-3' (sense) and 5'-CTCGATAATGGGGAAACAGTGT-3' (antisense). A detailed protocol for chromatin crosslinking, immunoprecipitation and PCR has been published³¹.

Received 24 March; accepted 2 August 2000.

- Norton, J. D., Deed, R. W., Craggs, G. & Sablitzky, F. Id helix-loop-helix proteins in cell growth and differentiation. *Trends Cell Biol.* **8**, 58–65 (1998).
- Massari, M. E. & Murre, C. Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. *Mol. Cell Biol.* **20**, 429–440 (2000).
- Iavarone, A., Garg, P., Lasorella, A., Hsu, J. & Israel, M. A. The helix-loop-helix protein Id-2 enhances cell proliferation and binds to the retinoblastoma protein. *Genes Dev.* **8**, 1270–1284 (1994).
- Lasorella, A., Iavarone, A. & Israel, M. A. Id2 specifically alters regulation of the cell cycle by tumor suppressor proteins. *Mol. Cell Biol.* **16**, 2570–2578 (1996).
- Clarke, A. R. *et al.* Requirement for a functional Rb-1 gene in murine development. *Nature* **359**, 328–330 (1992).
- Jacks, T. *et al.* Effects of an Rb mutation in the mouse. *Nature* **359**, 295–300 (1992).
- Lee, E. Y. -H. P. *et al.* Mice deficient for Rb are nonviable and show defects in neurogenesis and hematopoiesis. *Nature* **359**, 288–294 (1992).
- Zacksenhaus, E. *et al.* pRb controls proliferation, differentiation, and death of skeletal muscle cells and other lineages during embryogenesis. *Genes Dev.* **10**, 3051–3064 (1996).
- Lee, E. Y. *et al.* Dual roles of the retinoblastoma protein in cell cycle regulation and neuron differentiation. *Genes Dev.* **8**, 2008–2021 (1994).
- Jen, Y., Manova, K. & Benezra, R. Each member of the Id gene family exhibits a unique expression pattern in mouse gastrulation and neurogenesis. *Dev. Dyn.* **208**, 92–106 (1997).

- Neuman, T. *et al.* Neuronal expression of regulatory helix-loop-helix factor Id2 gene in mouse. *Dev. Biol.* **160**, 186–195 (1993).
- Zhu, W. *et al.* Id gene expression during development and molecular cloning of the human Id-1 gene. *Brain Res. Mol. Brain Res.* **30**, 312–326 (1995).
- Sherr, C. J. Cancer cell cycles. *Science* **274**, 1672–1677 (1996).
- Weinberg, R. A. The retinoblastoma protein and cell cycle control. *Cell* **81**, 323–330 (1995).
- Martinsen, B. J. & Bronner-Fraser, M. Neural crest specification regulated by the helix-loop-helix repressor Id2. *Science* **281**, 988–991 (1998).
- Maris, J. M. & Matthay, K. K. Molecular biology of neuroblastoma. *J. Clin. Oncol.* **17**, 2264–2279 (1999).
- Easton, J., Wei, T., Lahti, J. M. & Kidd, V. J. Disruption of the cyclin D/cyclin-dependent kinase/INK4/retinoblastoma protein regulatory pathway in human neuroblastoma. *Cancer Res.* **58**, 2624–2632 (1998).
- Beltinger, C. P., White, P. S., Sulman, E. P., Maris, J. M. & Brodeur, G. M. No CDKN2 mutations in neuroblastomas. *Cancer Res.* **55**, 2053–2055 (1995).
- Dicianni, M. B., Chau, L. S., Batova, A., Vu, T. Q. & Yu, A. L. The p16 and p18 tumor suppressor genes in neuroblastoma: implications for drug resistance. *Cancer Lett.* **104**, 183–192 (1996).
- Dicianni, M. B. *et al.* Frequent deregulation of p16 and the p16/G1 cell cycle-regulatory pathway in neuroblastoma. *Int. J. Cancer* **80**, 145–154 (1999).
- Castresana, J. S., Gomez, L., Garcia-Miguel, P., Queizan, A. & Pestana, A. Mutational analysis of the p16 gene in human neuroblastomas. *Mol. Carcinog.* **18**, 129–133 (1997).
- Kawamata, N., Seriu, T., Koeffler, H. P. & Bartram, C. R. Molecular analysis of the cyclin-dependent kinase inhibitor family: p16(CDKN2/MTS1/INK4A), p18(INK4C) and p27(Kip1) genes in neuroblastomas. *Cancer* **77**, 570–575 (1996).
- Fan, X., Gomez, L., Nistal, M., Sierrasesumaga, L. & Castresana, J. S. Lack of gene amplification as a mechanism of CDK4 activation in human neuroblastoma. *Oncol. Rep.* **6**, 647–650 (1999).
- Alevizopoulos, K., Vlach, J., Hennecke, S. & Amati, B. Cyclin E and c-Myc promote cell proliferation in the presence of p16INK4a and of hypophosphorylated retinoblastoma family proteins. *EMBO J.* **16**, 5322–5333 (1997).
- Goodrich, D. W. & Lee, W. H. Abrogation by c-Myc of G1 phase arrest induced by RB protein but not by p53 [published erratum: *Nature* **360**, 491 (1992)]. *Nature* **360**, 177–179 (1992).
- Amati, B., Alevizopoulos, K. & Vlach, J. Myc and the cell cycle. *Front Biosci.* **3**, D250–268 (1998).
- Slack, R. S., El-Bizri, H., Wong, J., Belliveau, D. J. & Miller, F. D. A critical temporal requirement for the retinoblastoma protein family during neuronal determination. *J. Cell Biol.* **140**, 1497–1509 (1998).
- Iavarone, A. & Massague, J. Repression of the CDK activator Cdc25A and cell-cycle arrest by cytokine TGF-beta in cells lacking the CDK inhibitor p15. *Nature* **387**, 417–422 (1997).
- Pietenpol, J. A. *et al.* TGF-β1 Inhibition of c-Myc transcription and growth in keratinocytes is abrogated by viral transforming protein with pRB binding domains. *Cell* **61**, 777–785 (1990).
- Neuman, K., Nornes, H. O. & Neuman, T. Helix-loop-helix transcription factors regulate Id2 gene promoter activity. *FEBS Lett.* **374**, 279–283 (1995).
- Boyd, K. E., Wells, J., Gutman, J., Bartley, S. M. & Farnham, P. J. c-Myc target gene specificity is determined by a post-DNA-binding mechanism. *Proc. Natl Acad. Sci. USA* **95**, 13887–13892 (1998).
- Morrow, M. A., Mayer, E. W., Perez, C. A., Adlam, M. & Siu, G. Overexpression of the Helix-Loop-Helix protein Id2 blocks T cell development at multiple stages. *Mol. Immunol.* **36**, 491–503 (1999).
- Florio, M. *et al.* Id2 promotes apoptosis by a novel mechanism independent of dimerization to basic helix-loop-helix factors. *Mol. Cell Biol.* **18**, 5435–5444 (1998).
- Cooper, C. L., Brady, G., Bilia, F., Iscove, N. N. & Quesenberry, P. J. Expression of the Id family helix-loop-helix regulators during growth and development in the hematopoietic system. *Blood* **89**, 3155–3165 (1997).
- Tsai, K. Y. *et al.* Mutation of E2f-1 suppresses apoptosis and inappropriate S phase entry and extends survival of Rb-deficient mouse embryos. *Mol. Cell* **2**, 293–304 (1998).
- Jen, Y., Manova, K. & Benezra, R. Expression patterns of Id1, Id2, and Id3 are highly related but distinct from that of Id4 during mouse embryogenesis. *Dev. Dyn.* **207**, 235–252 (1996).
- Yokota, Y. *et al.* Development of peripheral lymphoid organs and natural killer cells depends on the helix-loop-helix inhibitor Id2. *Nature* **397**, 702–706 (1999).
- Lufkin, T. *et al.* Homeotic transformation of the occipital bones of the skull by ectopic expression of a homeobox gene. *Nature* **359**, 835–841 (1992).
- Jacobs, J. J., Kieboom, K., Marino, S., DePinho, R. A. & van Lohuizen, M. The oncogene and Polycomb-group gene bmi-1 regulates cell proliferation and senescence through the ink4a locus. *Nature* **397**, 164–168 (1999).
- Serrano, M., Lin, A. W., McCurrach, M. E., Beach, D. & Lowe, S. W. Oncogenic ras provokes premature cell senescence associated with accumulation of p53 and p16INK4a. *Cell* **88**, 593–602 (1997).
- Iavarone, A. & Massague, J. E2F and histone deacetylase mediate transforming growth factor beta repression of cdc25A during keratinocyte cell cycle arrest. *Mol. Cell Biol.* **19**, 916–922 (1999).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank Y. Yokota and T. Jacks for providing us with *Id2*^{+/−} and *Rb*^{+/−} mice, respectively; R. Russell for support in the preparation of pathological specimens and histological consultation; J. Massague for the Mv1Lu cell line expressing tetracycline-inducible c-Myc; E. Latres for *p15*^{−/−} MEFs; C. Thiele for human neuroblastoma cell lines; R. Pestell for c-MycER cDNA; N. Schreiber-Agus for *N-myc* cDNA; M. van Lohuizen for LZRS-GFP retroviruses; and S. Lowe for retroviral vectors encoding activated *ras* and *E1A*. This work was supported by a grant from AIRC (Associazione Italiana per la Ricerca sul Cancro) to A.L., a grant from NIH (R01 to A.L.) and an institutional research grant from the American Cancer Society (A.L.). A.L. is recipient of a Sinsheimer Scholar award and M.N. is supported by a fellowship from Fondazione Beatrice Angelini.

Correspondence and requests for materials should be addressed to A.I. (e-mail: iavarone@acom.yu.edu). The sequence of human *Id2* promoter is deposited in GenBank under accession number AF270493.

Structural determinants of water permeation through aquaporin-1

Kazuyoshi Murata^{*†}, Kaoru Mitsuoka^{†‡}, Teruhisa Hirai^{†‡}, Thomas Walz[§], Peter Agre^{||}, J. Bernard Heymann[¶], Andreas Engel[¶] & Yoshinori Fujiyoshi[‡]

^{*} National Institute for Physiological Sciences, Okazaki 444-8585, Japan

[‡] Department of Biophysics, Faculty of Science, Kyoto University, Kitashirakawa, Sakyo-ku, Kyoto, 606-8502, Japan

[§] Department of Cell Biology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts, 02115-5730, USA

^{||} Departments of Biological Chemistry and Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205-2185, USA

[¶] M. E. Müller-Institute for Microscopy at the Biozentrum, University of Basel, Basel CH-4056, Switzerland

[†] These authors contributed equally to this work.

Human red cell AQP1 is the first functionally defined member of the aquaporin family of membrane water channels. Here we describe an atomic model of AQP1 at 3.8 Å resolution from electron crystallographic data. Multiple highly conserved amino-acid residues stabilize the novel fold of AQP1. The aqueous pathway is lined with conserved hydrophobic residues that permit rapid water transport, whereas the water selectivity is due to a constriction of the pore diameter to about 3 Å over a span of one residue. The atomic model provides a possible molecular explanation to a longstanding puzzle in physiology—how membranes can be freely permeable to water but impermeable to protons.

The cells of all life forms are surrounded by a membrane and filled with water. Although simple lipid bilayers exhibit limited water permeability, membranes of red cells, cells in renal proximal tubules and in certain other tissues are extremely permeable to water. Water-selective membrane channel proteins were predicted in these tissues to explain the high water permeability (low activation energy) and reversible inhibition by mercuric ions¹. AQP1 was first identified as an integral membrane protein in red blood cells and renal proximal tubules^{2,3}, where it functions as a water-selective membrane pore⁴. The aquaporin water channel family is found throughout nature and is involved in numerous physiological processes⁵.

Biological membranes exquisitely control the entry and exit of ions from cells. Aquaporins have to be highly specific for water to prevent other solutes and ions from also crossing the membrane. Protons present a particularly difficult challenge, because the positive charge of a proton (H_3O^+) can move along a column of water by hydrogen bond exchange⁶. Proton fluxes across cellular membranes drive physiological processes, such as membrane fusion, vesicular transport, solute transport and ATP synthesis.

The first insight into the structure and function of AQP1 came from sequence analyses and expression studies in *Xenopus laevis* oocytes. The AQP1 monomer contains 269 amino-acid residues³, which form two tandem repeats of three membrane-spanning α -helices with amino- (N-) and carboxy- (C-) termini located on the cytoplasmic side of the membrane⁷⁻⁹. In the hourglass model¹⁰, connecting loops B (cytoplasmic) and E (extracellular) each contain the consensus motif Asn-Pro-Ala and dip into the membrane from the opposite sides where they overlap, forming a single transmembrane aqueous pathway through each subunit of the AQP1 tetramer^{11,12}. Reconstitution of highly purified human red cell AQP1 into well-ordered membrane crystals has permitted definition of AQP1 structure at increasingly higher levels of resolution¹³⁻¹⁶. Here we present an atomic model of AQP1 and discuss the structural basis for water selectivity.

Density map from electron crystallography

The two-dimensional (2D) crystals of AQP1 used in this study were of limited size, typically 1 μm in diameter, making high-resolution data collection difficult especially at highly tilted conditions, whereas diffraction patterns from non-tilted crystals showed clear spots to a higher resolution than 3.5 Å. A helium-cooled electron

microscope^{17,18} helped us to improve the resolution of our structural analysis by reducing the effects of radiation damage as well as by enabling us to record a large number of electron micrographs and diffraction patterns. By selecting only images and electron diffraction patterns taken from well-ordered crystals, as judged by image analysis and comparison with a preliminary merged three-dimensional (3D) data set; we calculated a 3D density map at a resolution of 3.8 Å in the membrane plane and 4.6 Å normal to the membrane (Table 1).

The AQP1 monomer contains six tilted, membrane-spanning α -helices forming a right-handed bundle^{13,16,19}. The 3.8 Å map clearly resolves protrusions corresponding to individual side chains on membrane-spanning α -helices; these permitted unambiguous assignment of the α -helical structure. We established the quality and correctness of the final atomic model by alignment of 164 aquaporin sequences²⁰, computational fitting of α -helical fragments to the 4.5 Å map²¹, and by mutagenesis¹⁰.

The quality of our crystallographic data was relatively poor, as seen in the phase residual of 40.3° (Table 1). Nevertheless, the phase

Table 1 Electron crystallographic data

Two-dimensional crystals	
Layer group	p42 ₁ 2
Unit cell	$a = b = 96.0 \text{ Å}$
Thickness	100 Å (assumed)
Electron diffraction	
Number of diffraction patterns merged	135
Resolution	3.8 Å (in membrane plane) 4.4 Å (normal to membrane plane)
R_{Friedel}	44.0%
R_{merge}	49.5% (54.3% from 4.3 to 3.8 Å)
Observed amplitudes to 3.8 Å	85,254
Maximum tilt	61.1°
Fourier space sampled	88.0%
Phase determination by image processing	
Processed images	103
Resolution	3.8 Å (in membrane plane) 4.6 Å (normal to membrane plane)
Average phase residual	40.3° (58.2° from 4.3 to 3.8 Å)
Observed phases to 3.8 Å	24,181
Maximum tilt	62.0°
Fourier space sampled*	54.0%
Crystallographic refinement	
Crystallographic R factor (96 Å to 3.8 Å)	39.9%
Free R factor (96 Å to 3.8 Å)	41.7%

* Only reflections with a figure of merit over 0.5 were included.

residual corresponds to a figure of merit of 0.76, similar or better than the initial phases ordinarily determined by X-ray crystallography. The quality of the map was sufficient for model building (Fig. 1), although surface loops were less well defined, so their positions represent the most likely assignment.

Protein fold of the AQP1 monomer

The order of α -helices in the AQP1 monomer was previously

unknown. Beginning at the four-fold axis of the tetramer, the helices are arranged: 2-1-3 (first repeat), 5-4-6 (second repeat). In the first repeat, helices 1 and 2 run almost parallel to each other, and are tilted to approximately the same angle. Helix 3 is oriented almost perpendicular to the axis roughly defined by the top-on view of the first two helices. The arrangement of the first and second repeat exhibits a pseudo two-fold axis running parallel to the membrane plane at the centre of the molecule (Fig. 2a-c).

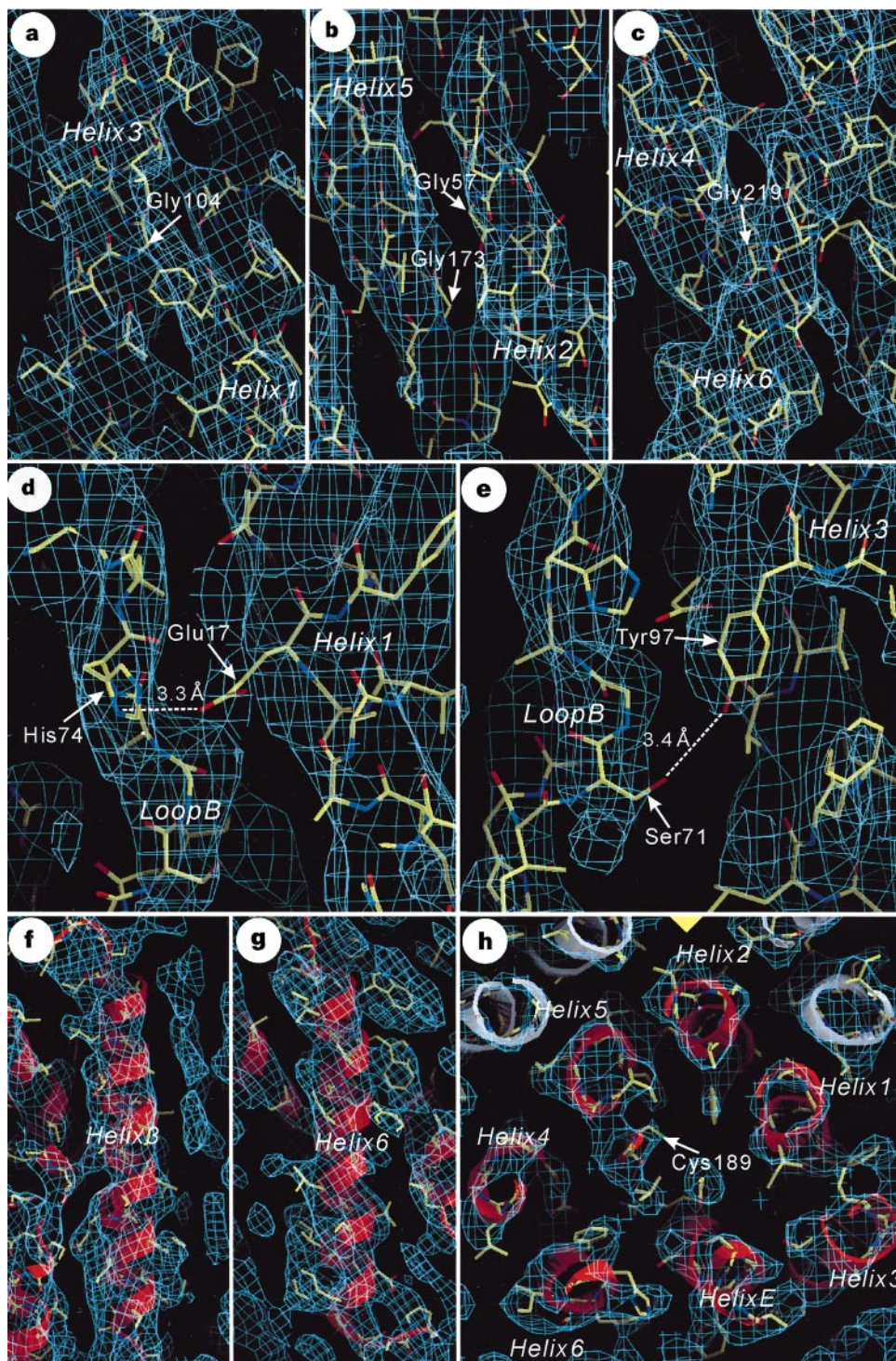


Figure 1 Representative areas of the AQP1 potential map with the fitted atomic model. **a**, Interaction of helices 1 and 3; the arrow depicts Gly 104. **b**, Ridges into grooves structure for the stable interaction between helices 2 and 5; the arrows show Gly 57 and Gly 173. **c**, Interaction of helices 4 and 6; the arrow indicates Gly 219. **d**, Ion pair formed by Glu 17 and His 74. **e**, Hydrogen bonding between Ser 71 and Tyr 97 (one of the noisiest

parts of the potential map at the membrane surface). **f**, **g** Density for helix 3 and the corresponding helix 6. A cluster of large protrusions were assigned to aromatic residues Trp 210, Phe 212, and Trp 213 in helix 6; no bulky protrusions are found in the density assigned for helix 3. **h**, Section through the monomer close to Cys 189 in loop E. The yellow triangle indicates the four-fold axis of the AQP1 tetramer.

The short N terminus of AQP1 is located on the cytoplasmic side of the membrane (Fig. 3). Helix 1 then crosses the membrane and is connected to helix 2 by the extended extracellular loop A with the glycosylation site Asn 41 forming a weak density in our map. Helix 2 returns back through the membrane adjacent to helix 1 near the four-fold axis of the AQP1 tetramer. At the cytoplasmic end of helix 2, loop B folds back into the membrane positioning the Asn-Pro-Ala motif midway through the membrane. Following the Asn-Pro-Ala motif, loop B forms the short pore helix HB located adjacent to helix 6. At the cytoplasmic surface farthest from the four-fold axis, HB emerges from the membrane and connects to helix 3, which crosses the membrane adjacent to helix 1 and termi-

nates on the extracellular surface. The first and second repeat of each monomer are connected by the 18-residue loop C which spans the entire extracellular surface of the monomer joining helix 3 to helix 4 (Fig. 2a and b). Oriented opposite to the first repeat, the second repeat begins at the extracellular surface as helix 4 which crosses the membrane and is connected to helix 5 by the short loop D near the four-fold axis. Helix 5 returns back through the membrane between helix 4 and helix 2 to the extracellular surface. Loop E folds back into the membrane from the extracellular side, and the Asn-Pro-Ala motif is positioned near the centre of the molecule where it crosses the Asn-Pro-Ala motif in loop B at a right angle (Fig. 2c, 4a and c), and forms the short pore helix HE. Helix 6 crosses the membrane adjacent to



Figure 2 Ribbon representations of the AQP1 fold depicting the six membrane-spanning helices, the two pore helices, and the connecting loops in different colours. **a**, End-on view from the extracellular surface; **b**, side view; **c**, end-on view from the cytoplasmic surface. **d**, Cylinder model of the AQP1 tetramer; end-on view from the extracellular surface;

e, side view. Yellow diamonds and yellow dotted line indicate the four-fold axis of the AQP1 tetramer. Dotted lines and spindle in white show the pseudo two-fold axis of the AQP1 monomer. Grey bands indicate the surface of the lipid bilayer. **a–c** were prepared with MOLSCRIPT⁴⁵ and Raster 3D⁴⁶.

helix 4, with the C-terminus emerging on the cytoplasmic side of the membrane (Fig. 3).

Three large helical crossing angles stabilize the right-handed coiled-coil interactions (Table 2). Helices 1 and 3 (36.7°), helices 4 and 6 (40.9°), and helices 2 and 5 (28.5°) all have highly conserved glycine residues at the contact sites. The glycine residues at the contact region, Gly 104 for helices 1 and 3, Gly 219 for helices 4 and 6, are shown in Fig. 1a and 1c. The crossing of helix 2 (first repeat) and helix 5 (second repeat) is maintained by interactions involving highly conserved residues Gly 57 and Gly 173 (Fig. 1b). The fitting of ridges into grooves in helices 2 and 5 lock the two AQP1 helical bundles together closely at the four-fold axis of the tetramer, opposite to the site where HB and HE interact with helices 6 and 3 (Fig. 2a). The conserved glycines²⁰ in helices 3 and 6 reveal the GlyxxxGly motif that is a framework for transmembrane helix-helix association²² (Ala 100, Gly 104, Ala 108 in helix 3, Gly 215, Gly 219, Ala 223 in helix 6; glycine can sometimes be replaced by alanine in the motif). Thus, the α -helices in both repeats are held together near the interior and exterior surfaces of the monomer. Gly 82 in pore helix HB forms the contact site with helix 6 and Gly 198 in pore helix HE interacts with helix 3.

Additional highly conserved residues are concentrated in and around the two functional loops B and E that are closely positioned in the interior of the monomer²⁰. Loops B and E are held together by van der Waals interactions between the prolines in the two Asn-Pro-Ala motifs (Pro 77 and Pro 193) that cross each other at an angle of approximately 90° (Fig. 4a and c). The asparagines in the Asn-Pro-Ala motifs cap the N-terminal end of the pore helices by hydrogen bonding with the main-chain NH groups at Val 79 and Arg 195. The positions of functional loops B and E are stabilized through ion pairs and hydrogen bonds. His 74 (loop B) forms an ion pair with the conserved Glu 17 (helix 1) anchoring the position of pore loop B (Fig. 1d), whereas Arg 195 in pore helix HE is connected by a salt bridge to the conserved Glu 142 (helix 4). The density for Arg 195 is less clear and could be pointing towards the membrane surface; however, this possibility seems to be very unlikely because of the conservation of this arginine in the aquaporin family. These ion pairs are stabilized by conserved polar residues that surround them within the membrane, Thr 80, Gln 101 and Ser 196. The positions of the NH groups in the main chain of the pore loops are stabilized by hydrogen bonds to highly conserved threonine residues, Thr 21

(helix 1) and Thr 146 (helix 4). In addition, the side chain of Ser 71 (loop B) forms a hydrogen bond with Tyr 97 (helix 3), further stabilizing the position of this loop (Fig. 1e).

AQP1 tetramer

AQP1 is a homotetramer (Fig. 2d and e). Stability may result from accommodation of each monomer as a tight-fitting wedge within the tetramer. The relative insolubility of AQP1 in certain detergents¹² indicates that surrounding lipids may be tightly attached. Sufficient space lies between adjacent tetramers in the 2D membrane crystals to accommodate at least one lipid molecule that may be important in crystal formation. Amino-acid residues interacting with the acyl chains of the lipids in the tetramer are found in helices 3 and 6 and pore helices HB and HE, which reside at the exteriors of the monomers.

Each monomer interacts with two neighbouring monomers through membrane-spanning α -helices. Helices 1 and 2 form left-handed coiled-coil interactions with helices 4 and 5 of the neighbouring monomer with crossing-angles of 48.5° and 48.0° . Helix 1 also interacts with helix 5 of the adjacent monomer at the extracellular surface, whereas helix 2 interacts with helix 4 of the adjacent monomer at the cytoplasmic surface. These interactions are probably stabilized by a network of hydrogen bonds between residues Ser 59, Thr 62, and Gln 65 (helix 2) with residues Gln 148, Cys 152, and Thr 156 (helix 4 of the adjacent monomer).

Interactions between loops may also contribute to tetramer stability. Loops A in the four monomers surround the four-fold axis of the tetramer on the extracellular surface, whereas loops D surround the axis on the cytoplasmic surface. The 3.8-Å structure

Table 2

	Start	End	Tilt angle to membrane normal (degree)	Angle between two helices (degree)
Helix 2	49	66	23.7	17.5 (h2 to h1)
Helix 1	8	36	27.2	36.7 (h1 to h3)
Helix 3	95	115	39.9	9.0 (h3 to hB)
Helix B	77	84	34.2	33.9 (hB to h6)
				42.4 (hE to h3)
Helix E	193	200	37.6	4.9 (hE to h6)
Helix 6	208	228	33.5	40.9 (h6 to h4)
Helix 4	137	155	24.3	5.8 (h4 to h5)
Helix 5	167	183	24.8	28.5 (h2 to h5)

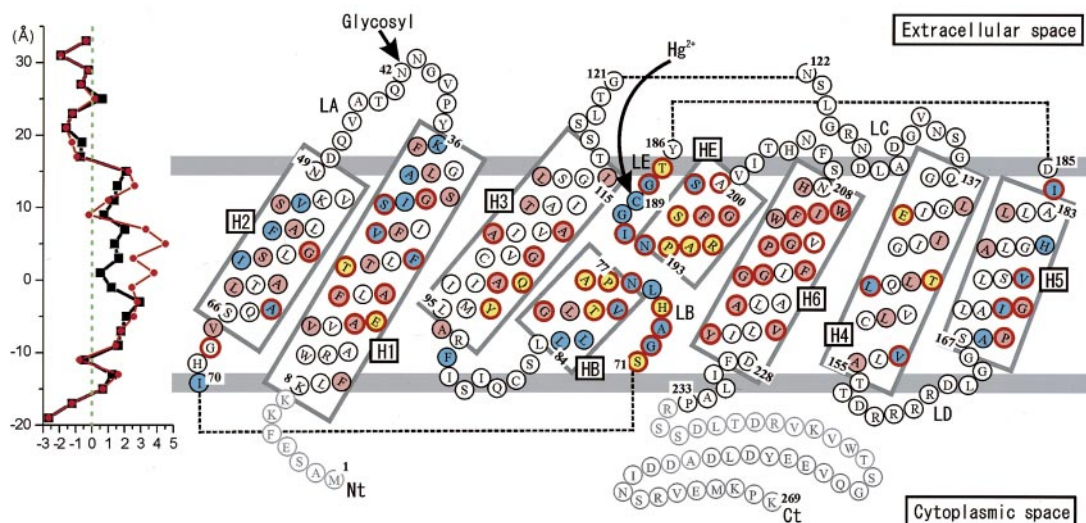


Figure 3 Schematic representation of AQP1 topology and hydropathy. Secondary structural elements in AQP1 indicating structurally important amino-acid residues: those that face the inside of the aqueous pore in the membrane (blue); highly conserved residues (thick red circles); residues stabilizing loop-associations (yellow); contact points for helix-helix interaction (pink). The surfaces of the lipid bilayer are represented by grey

bands. The density map did not allow us to trace the seven N- and 36 C-terminal amino-acid residues (grey). On the left side, the hydropathy of the molecular surface of AQP1 is shown, indicating the boundaries of the lipid bilayer. Hydropathy plots were calculated from all amino-acid residues (red) and from residues not buried (black).

reveals a protrusion from the extracellular surface of the AQP1 tetramer, whereas the cytoplasmic side forms a shallow depression (Fig. 2b and e). The N terminus extends from the periphery of the tetramer on the cytoplasmic side in close proximity to the C terminus of the adjacent monomer.

Helices 2 and 5 incline towards the four-fold axis at the extra-cellular side and the cytoplasmic side of the membrane (Fig. 2d). The narrowest parts about the four-fold axis reside close to both membrane surfaces near Val 50 (helix 2) and Leu 170 (helix 5). At this level, the gap in the space-filling model without hydrogen atoms has a diameter of about 3.5 Å. At its widest, the diameter of the gap is approximately 8.5 Å, which is insufficient to accommodate a lipid molecule. The hydrophobic environment, however, suggests that this space is filled with a hydrophobic molecule. Thus, the space at the four-fold axis of the tetramer does not form a pathway for water.

Structure of the aqueous pore

The critical function of AQP1 is its exceptional water permeability, 3×10^9 molecules per monomer per second^{23–25}. As predicted by the hourglass model¹⁰, loops B and E interact with each other through

their Asn-Pro-Ala motifs, forming part of the surface of the aqueous pore (Fig. 4a and b). Helices 2 and 5 together with the C-terminal halves of helices 1 and 4 form the remaining surface of the pore (blue residues, Fig. 3). The 3 Å diameter of the narrowest part of the pore measured by van der Waals distance, determined from a carbon–carbon distance of 6.5 Å, is only slightly larger than the 2.8 Å diameter of a water molecule (Fig. 4a). Although the span of this constriction is only one amino-acid residue, this would block passage of larger solutes and ions. Unlike the potassium channel²⁶, AQP1 does not contain a structure to liberate ions from their hydration shell. Thus, the atomic model of AQP1 is consistent with biophysical measurements that showed no solute, ion or proton permeability^{25,27,28}.

The 3 Å constriction in the pore is located in the centre of the membrane just beside the region where loops B and E interact with each other. Ile 60 (helix 2), Phe 24 (helix 1), Leu 149 (helix 4) and Val 176 (helix 5) form a hydrophobic surface lining the inside of the pore adjacent to Asn 76 and Asn 192 of the Asn-Pro-Ala motifs (Fig. 4a and b). Two histidines are located within 5 Å of the pore constriction. His 74 is highly conserved and forms an ion pair with Glu 17 (Fig. 1d). His 180 is conserved in water-selective

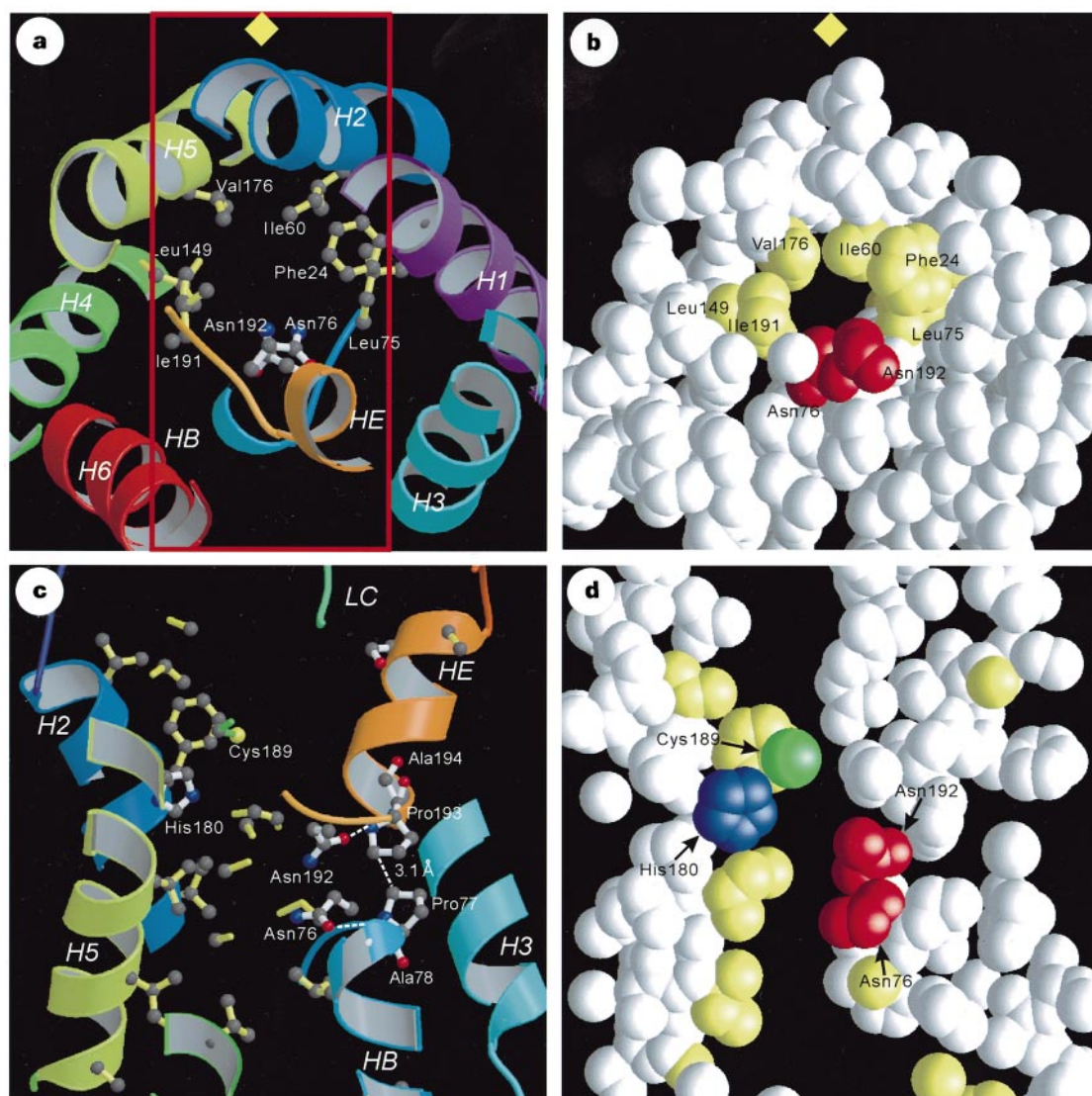


Figure 4 Ribbon and space-filled diagrams representing the structure of the aqueous pore in AQP1. **a**, Section cut parallel to the plane of the membrane through the centre of the AQP1 monomer showing the constriction of the pore. **b**, Space-filled model of the same view as **a** showing the pore diameter. **c**, Section cut across the membrane as outlined by red rectangle-lines perpendicular to the plane in **a**, showing the side view of the interior of the AQP1 pore. **d**, Space-filled model of the same view as in panel **c**

showing the widening (top and bottom) and constriction of the pore (centre). Yellow diamonds in **a** and **b** show the four-fold axis of the AQP1 tetramer. Side chains facing the inside of the aqueous pore are indicated: hydrophobic side chains at the pore constriction (yellow); Asn 76 and Asn 192 (red); cysteine (green); His 180 (blue). **a–d** were prepared with MOLSCRIPT⁴⁵ and Raster 3D⁴⁶.

members of the aquaporin family but is usually replaced by glycine in homologues permeated by glycerol²⁰, suggesting that it plays a specific role in water conductance. Hydrophobicity of the inner pore surface and small axial dimension (5 Å) of the constriction are probably of crucial importance for facilitating the exceptionally high water permeability of AQP1.

The free sulphhydryl of Cys 189 in loop E projects into the pore from the extracellular side of the narrowing (Figs 1h, 3, 4c and d) where inhibition of AQP1 water permeation by mercury^{29,30} must result from occlusion of the aqueous pathway. As expected for a symmetrical structure, Ala 73, the corresponding residue in loop B, resides at a comparable location on the cytoplasmic side of the narrowing, explaining the inhibition of the double mutant Cys 189 Ser, Ala 73 Cys by mercury¹⁰.

Permeation by water but not protons

As studied in detail with gramicidin A, water can readily cross cell membrane as a continuous unbroken column of molecules within an open pore. This single-file hydrogen-bonded chain of water molecules conducts protons with great efficiency⁶. In contrast, no proton conductance was detected in AQP1 (ref. 25), so the pore must prevent their passage. The mechanism that prevents transfer of protons during rapid water permeation remains an important unanswered question in the structural biology of aquaporins. Blocking proton transfer is believed to require interruption of the continuous chain of hydrogen bonds along a single file of water by hydrogen-binding sites at the pore surface⁶. The hydrogen-bonding energy is about 3 kcal (ref. 31), close to the activation energy for the passage of water through AQP1 (ref. 24), suggesting that breakages of hydrogen bonds are involved in the water flow.

The pore helices HB and HE and the neighbouring Asn residues in the two Asn-Pro-Ala motifs may be critical elements for this process. Short pore helices were also identified in the atomic structure of the potassium channel from *Streptomyces lividans*²⁶. The C-termini of those pore helices point into the large aqueous cavity surrounded by the membrane-spanning helices. In contrast, the AQP1 pore helices are oriented in the opposite direction, the C-termini facing out. Due to the positive electrostatic field generated by the dipole moments of the pore helices in AQP1, the oxygen atom in a water molecule coming close to the membrane centre orients to the side of the Asn-Pro-Ala motif (Fig. 5a). We note that the shape of the AQP1 pore is the opposite of the pore shape formed by the potassium channel²⁶, having a constriction in the centre of the membrane and wide openings at the membrane surfaces. This constriction results in a high dielectric barrier that repels ions, while allowing penetration of neutral solutes.

The pore helices HB and HE are held in the middle of the membrane by van der Waals interactions between the Pro residues of the two Asn-Pro-Ala motifs. Our atomic model suggests hydrogen bonds between the main chain NH groups and the carbonyl groups of the two asparagine residues in the Asn-Pro-Ala motifs, thus capping the N-terminal end of the short pore helices. As a result, the amido groups of Asn 76 and 192 are fixed to extend into the pore at the constriction (Fig. 5). The oxygen atom of the water molecule here will form hydrogen bonds with these amido groups by changing the hydrogen-bonding partner from adjacent water molecules (Fig. 5b). This reorients the two hydrogen atoms of the water molecule at the pore constriction perpendicular to the channel axis because of the arrangement of the molecular orbital for water (Fig. 5c). Thus, the two hydrogen atoms of the water molecule are prevented from forming hydrogen bonds with adjacent water molecules in the single-file column. Because all residues near the constriction are hydrophobic, no hydrogen-bonding partners are provided by the opposite wall of the pore, giving a maximum energy cost of 3 kcal, because only one hydrogen bond relative to bulk water is affected. The water molecule in the pore constriction can form hydrogen bonds via oxygen but not through

hydrogen atoms. Thus, water molecules can permeate the pore with a minimal energy barrier, whereas transfer of protons is blocked by hydrogen-bond isolation from bulk water (H-bond isolation mechanism). Because our analyses were restricted to the resolution of 3.8 Å, no water molecules could be observed. Higher-resolution structural studies may prove our hypothesis if the occupancy of water molecules at the pore constriction is sufficiently high.

The aquaporin family

Although our atomic model explains several key features of the AQP1 pore, many open questions remain concerning other members of the aquaporin family. For example, how can the same protein fold account for the extreme water-selectivity of AQP1 and the broad substrate specificity of AQP9 (ref. 32)? How can aquaglyceroporins allow passage of glycerol but be less permeable to water than AQP1? How do anions pass through AQP6 (ref. 33)? What are the molecular mechanisms for regulation of several aquaporins?

The atomic model provides a precedent for modelling other aquaporins by carefully threading their amino-acid sequences along the main-chain fold of AQP1. The models can then be used to make predictions about which residues are important for pore selectivity. Site-directed mutagenesis of key residues in the atomic model of AQP1 will permit assessment of their functional importance. This should open new avenues for investigating the structural determinants of pore specificities in other members of the aquaporin family. Although low-resolution density maps of other aquaporins will provide guidance for modelling, the atomic structures are needed to bring unambiguous identification of the molecular determinants.

Note added in proof: The structure of the glycerol facilitator (GlpF), an aquaglyceroporin from *E. coli*, has now been determined⁴⁷. □

Methods

Data collection

Two-dimensional crystals of AQP1 were produced as described previously²³, and prepared

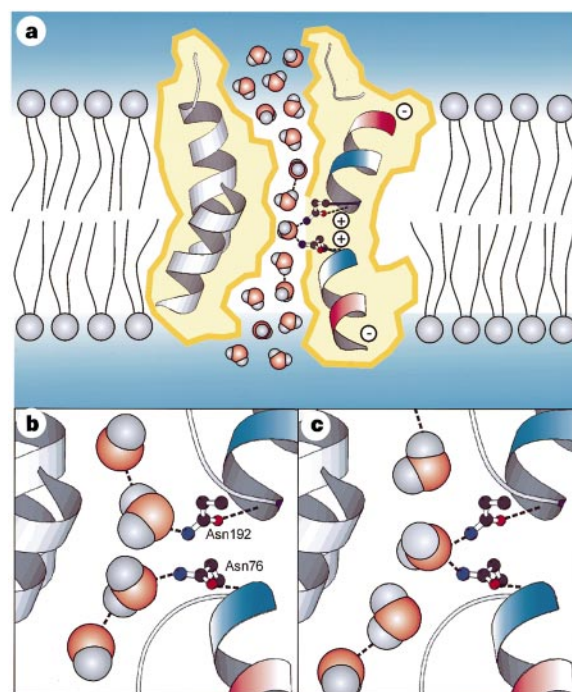


Figure 5 Schematic representations explaining the mechanism for blocking proton permeation of AQP1. **a**, Diagram illustrating how partial charges from the helix dipoles restrict the orientation of the water molecules passing through the constriction of the pore. **b** and **c** Diagram illustrating hydrogen bonding of a water molecule with Asn 76 and/or Asn 192, which extend their amido groups into the constriction of the pore. The ribbon and ball-stick models in this figure were prepared with MOLSCRIPT⁴⁵.

for electron cryomicroscopy on thin carbon-coated molybdenum grids using trehalose as embedding medium at a final concentration of 3%^{16,34}. Electron micrographs and diffraction patterns were recorded with a JEM 3000SFF electron microscope equipped with a liquid-helium stage^{17,18} and a field-emission electron source using an acceleration voltage of 300 kV. Electron micrographs recorded under low-dose conditions ($< 20 \text{ e}^- \text{Å}^{-2}$) were digitized with a LeafScan45 scanner (Scitex)³⁵ for image processing; electron-diffraction patterns were collected with a 2K slow-scan charge-coupled device (CCD) camera (Gatan)³⁶. Electron-diffraction patterns and images were processed with a modified version of the MRC image processing programs³⁷. Electron-diffraction patterns (173) tilted up to 61.1° were analysed to calculate amplitudes using an additional background-correction step³⁸. Images were computationally unbent and corrected for the contrast-transfer function³⁹. Images (124) tilted up to 62.0° were averaged to generate a merged-phase data set. The final three-dimensional potential map was produced with the CCP4 programs⁴⁰ from merged amplitudes and phases with figures of merit above 0.5.

A problem for the data collection arose from the relatively small size and lattice defects of the AQP1 2D crystals that prevented collection of data of very high resolution. Substantially more data were collected and processed than needed for calculation of a 3D density map at a resolution of 3.8 Å. Images and diffraction patterns were excluded from the final analysis if they did not significantly improve the statistics. Thus, our final data set includes 135 electron diffraction patterns (providing structure factor amplitudes) and 103 images (providing experimental phases) collected from specimens at tilt angles up to 60° . Statistics of our electron-crystallographic analysis of the AQP1 structure are summarized in Table 1.

Model building

The atomic model was built using the program O (ref. 41). Conventionally constrained (positional) refinement of the initial hand-built model was performed with the program X-PLOR⁴² without simulated annealing, with phase and amplitude constraints. The moderate resolution and the relatively poor quality of our electron-crystallographic data (Table 1) added to the challenge of building an atomic model for AQP1. The map is of a quality similar to initial non-refined electron-density maps determined by X-ray crystallography of proteins with large relative molecular mass. Furthermore, most of the densities in the membrane region of the 3D map were clearly defined. Only the extramembranous regions of the proteins contained some persistent ambiguities.

Each tandem repeat of AQP1 contains three membrane-spanning α -helices. The second half of the protein (helices 4–6) is incorporated into the membrane in an orientation opposite to the first half of the protein (helices 1–3). Thus, the two halves of the protein are related to each other by a pseudo two-fold axis that runs parallel to the membrane plane¹³ (Fig. 2a–c). This symmetry of the AQP1 structure constituted a major difficulty during the actual model building, because each density in the 3D map could be assigned to either of two possible helices in the AQP1 molecule—a helix in the first tandem repeat or the corresponding helix in the second tandem repeat. Of note, the densities for helices 3 and 6 were clearly distinguishable, because helix 6 contains a cluster of aromatic residues (Phe 206, His 209, Trp 210, Phe 212 and Trp 213) not present in helix 3 (Fig. 1f and g), whereas the accurate assignment of individual residues was difficult. Similarly, several bulky side chains in pore helix HB that are missing from pore helix HE enabled us to distinguish the two tandem repeats and provided a starting point for the model building. Our assignment is also supported by the sidedness of AQP1 and AqpZ (ref. 43) determined by atomic force microscopy⁴⁴.

PROCHECK was used to calculate a Ramachandran plot for our atomic model of residues 8–233 (missing seven N-terminal and 36 C-terminal amino-acid residues). 70 per cent of residues fell into the most favourable locations, whereas only five amino-acid residues from connecting-loop domains fell into the generously allowed regions.

Received 21 June; accepted 24 September 2000.

- Macey, R. I. & Farmer, R. E. I. Inhibition of water and solute permeability in human red cells. *Biochem. Biophys. Acta* **211**, 104–106 (1970).
- Denker, B. M., Smith, B. L., Kuhajda, F. P. & Agre, P. Identification, purification, and partial characterization of a novel Mr 28,000 integral membrane protein from erythrocytes and renal tubules. *J. Biol. Chem.* **263**, 15634–15642 (1988).
- Preston, G. M. & Agre, P. Isolation of the cDNA for erythrocyte integral membrane protein of 28 kilodaltons: Member of an ancient channel family. *Proc. Natl Acad. Sci. USA* **88**, 11110–11114 (1991).
- Preston, G. M., Carroll, T. P., Guggino, W. B. & Agre, P. Appearance of water channels in *Xenopus* oocytes expressing red cell CHIP28 protein. *Science* **256**, 385–387 (1992).
- Heymann, J. B. & Engel, A. Aquaporins: Phylogeny, structure and physiology of water channels. *News Physiol. Sci.* **14**, 187–193 (1999).
- Pomes, R. & Roux, B. Structure and dynamics of a proton wire: a theoretical study of H^+ translocation along the single-file water chain in the gramicidin A channel. *Biophys. J.* **71**, 19–39 (1996).
- Wistow, G., Pisano, M. & Chepelinsky, A. Tandem sequence repeats in transmembrane channel proteins. *Trends Biochem. Sci.* **16**, 170–171 (1991).
- Reizer, J., Reizer, A. & Saier, M. J. The MIP family of integral membrane channel proteins: sequence comparisons, evolutionary relationships, reconstituted pathway of evolution, and proposed functional differentiation of the two repeated halves of the proteins. *Crit. Rev. Biochem. Mol. Biol.* **28**, 235–257 (1993).
- Preston, G. M., Jung, J. S., Guggino, W. B. & Agre, P. Membrane topology of aquaporin CHIP. Analysis of functional epitope-scanning mutants by vectorial proteolysis. *J. Biol. Chem.* **269**, 1668–1673 (1994).
- Jung, J., Preston, G. M., Smith, B., Guggino, W. & Agre, P. Molecular structure of the water channel through aquaporin CHIP. The hourglass model. *J. Biol. Chem.* **269**, 14648–14654 (1994).
- van Hoek, A. N. *et al.* Functional unit of 30 kDa for proximal tubule water channels as revealed by radiation inactivation. *J. Biol. Chem.* **266**, 16633–16635 (1991).
- Smith, B. L. & Agre, P. Erythrocyte Mr 28,000 transmembrane protein exists as a multisubunit oligomer similar to channel proteins. *J. Biol. Chem.* **266**, 6407–6415 (1991).

- Cheng, A., van Hoek, A. N., Yeager, M., Verkman, A. S. & Mitra, A. K. Three-dimensional organization of a human water channel. *Nature* **387**, 627–630 (1997).
- Li, H., Lee, S. & Jap, B. K. Molecular design of aquaporin-1 water channel as revealed by electron crystallography. *Nature Struct. Biol.* **4**, 263–265 (1997).
- Walz, T., Smith, B., Agre, P. & Engel, A. The three-dimensional structure of human erythrocyte aquaporin CHIP. *EMBO J.* **13**, 2985–2993 (1994).
- Walz, T. *et al.* The three-dimensional structure of aquaporin-1. *Nature* **387**, 624–627 (1997).
- Fujiyoshi, Y. *et al.* Development of a superfluid helium stage for high-resolution electron microscopy. *Ultramicroscopy* **38**, 241–251 (1991).
- Fujiyoshi, Y. The structural study of membrane proteins by electron crystallography. *Adv. Biophys.* **35**, 25–80 (1998).
- Mitsuoka, K. *et al.* The structure of aquaporin-1 at 4.5 Å resolution reveals short α -helices in the center of the monomer. *J. Struct. Biol.* **128**, 34–43 (1999).
- Heymann, J. B. & Engel, A. Structural clues in the sequences of the aquaporins. *J. Mol. Biol.* **295**, 1039–1053 (2000).
- De Groot, B. L. *et al.* The fold of human aquaporin 1. *J. Mol. Biol.* (in the press).
- Russ, W. P. & Engelman, M. The GxxxG motif: a framework for transmembrane helix-helix association. *J. Mol. Biol.* **296**, 911–919 (2000).
- Walz, T., Smith, B., Zeidel, M., Engel, A. & Agre, P. Biologically active two-dimensional crystals of aquaporin CHIP. *J. Biol. Chem.* **267**, 1583–1586 (1994).
- Zeidel, M. *et al.* Reconstitution of functional water channels in liposomes containing purified red cell CHIP28 protein. *Biochemistry* **31**, 7436–7440 (1992).
- Zeidel, M. *et al.* Ultrastructure, pharmacologic inhibition, and transport selectivity of aquaporin channel-forming integral protein in proteoliposomes. *Biochemistry* **33**, 1606–1615 (1994).
- Doyle, D. A. *et al.* The structure of the potassium channel: Molecular basis of K^+ conduction and selectivity. *Science* **280**, 69–77 (1998).
- Gutierrez, A. M., Gonzales, E., Echevarria, M., Hernandez, C. S. & Whitembury, G. The proximal straight tubule (PST) basolateral cell membrane water channel: Selectivity characteristics. *J. Membr. Biol.* **143**, 189–197 (1995).
- Yang, B., van Hoek, A. N. & Verkman, A. S. Very high single channel water permeability of aquaporin-4 in baculovirus-infected insect cells and liposomes reconstituted with purified aquaporin-4. *Biochemistry* **36**, 7625–7632 (1997).
- Preston, G., Jung, J., Guggino, W. & Agre, P. The mercury-sensitive residue at cysteine 189 in the CHIP28 water channel. *J. Biol. Chem.* **268**, 17–20 (1993).
- Zhang, R., van Hoek, A. N., Biwersi, J. & Verkman, A. S. A point mutation at cysteine 189 blocks the water permeability of rat kidney water channel CHIP28. *Biochemistry* **32**, 2938–2941 (1993).
- Schulz, G. E. & Schermer, R. H. in *Principles of Protein Structure* (ed. Cantor, C. R.) 17–26 (Springer, New York, 1979).
- Tukaguchi, H. *et al.* Molecular characterization of a broad selectivity neural solute channel. *J. Biol. Chem.* **273**, 24737–24743 (1998).
- Yasui, M. *et al.* Rapid gating and anion permeability of an intracellular aquaporin. *Nature* **402**, 184–187 (1999).
- Hirai, T. *et al.* Trehalose embedding technique for high-resolution electron crystallography: application to structural study on bacteriorhodopsin. *J. Elec. Microsc.* **48**, 653–658 (1999).
- Mitsuoka, K., Murata, K., Kimura, Y., Namba, K. & Fujiyoshi, Y. Examination of the Leafscan 45, a line-illuminating micro-densitometer, for its use in electron crystallography. *Ultramicroscopy* **68**, 109–121 (1997).
- Krivanek, O. L. & Mooney, P. E. Applications of slow-scan CCD camera in transmission electron microscopy. *Ultramicroscopy* **49**, 95–108 (1993).
- Crowther, R. A., Henderson, R. & Smith, J. M. MRC image processing programs. *J. Struct. Biol.* **116**, 86–92 (1996).
- Mitsuoka, K. *et al.* The structure of bacteriorhodopsin at 3.0 Å resolution based on electron crystallography: Implication on the charge distribution. *J. Mol. Biol.* **286**, 861–882 (1999).
- Henderson, R., Baldwin, J. M., Downing, K. H., Lepault, J. & Zemlin, F. Structure of purple membrane from *Halobacterium halobium*: recording, measurement and evaluation of electron micrographs at 3.5 Å resolution. *Ultramicroscopy* **19**, 147–178 (1986).
- Collaborative Computational Project No. 4. The CCP4 Suite: Programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
- Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps. *Acta Crystallogr. A* **47**, 110–119 (1991).
- Brünger, A. T. X-PLOR Version 3.1-A System for X-ray Crystallography (Yale Univ. Press, New Haven/London, 1988).
- Scheuring, S. *et al.* The aquaporin sidedness revisited. *J. Mol. Biol.* **295**, 1271–1278 (2000).
- Scheuring, S. *et al.* High resolution AFM topographs of the *Escherichia coli* water channel aquaporin Z. *EMBO J.* **18**, 4981–4987 (1999).
- Kraulis, P. J. MOLSCRIPT—a program to produce both detailed and schematic plots of proteins structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
- Merritt, E. A. & Bacon, D. J. Raster 3D: photorealistic molecular graphics. *Methods Enzymol.* **277**, 505–524 (1997).
- Fu, D. X. *et al.* Structure of a glycerol conducting channel and the basis for its selectivity. *Science* (in the press).

Acknowledgements

The authors thank B. L. Smith for technical assistance. These studies were partially carried out in the International Institute for Advanced Research, Matsushita Electric Industrial Co., Ltd. The research was supported by grants from the Japan Society for the Promotion of Science (JSPS), US National Institutes of Health, the Maurice E. Müller Foundation, the Swiss National Foundation for Scientific Research, and the European Union.

Correspondence and requests for materials should be addressed to Y.F. (e-mail: yoshi@em.biophys.kyoto-u.ac.jp). The coordinates of AQP1 reported in this paper have been deposited in the Protein Data Bank (Research Collaboratory for Structural Bioinformatics), PDB ID no. 1FQY.



Efficient delivery of meteorites to the Earth from a wide range of asteroid parent bodies

D. Vokrouhlický* & P. Farinella†

* Institute of Astronomy, Charles University, Holešovičkách 2, 18000 Prague, Czech Republic

† Dipartimento di Astronomia, Università da Trieste, Trieste, Italy

Almost all meteorites come from asteroids, but identifying their specific parent bodies, and modelling their transport to the Earth, has proved to be difficult^{1,2}. The usual model^{1,3,4} of delivery through orbital resonances with the major planets^{5,6} has recently been shown^{7–10} to deplete the supply of meteorites much too rapidly to explain either the observed flux at the Earth, or the length of time the meteorites have spent in space (as measured by cosmic-ray exposure ages). Independently, it has been found that a force arising from anisotropically emitted thermal radiation from asteroidal fragments (the ‘Yarkovsky effect’) influences the fragments’ orbits in important ways^{11–14}. Here we report the results of a detailed model for the transport of meteorites to the Earth, which includes the Yarkovsky effect and collisional evolution of the asteroidal fragments. We find that the Yarkovsky effect significantly increases the efficiency of the delivery of meteorites to the Earth, while at the same time allowing a much wider range of asteroids to contribute to the flux of meteorites. Our model also reproduces the observed distribution^{15,16} of cosmic-ray exposure ages of stony meteorites.

The Yarkovsky effect, a recoil force due to anisotropically emitted thermal radiation, strongly affects the long-term orbit evolution of Solar System bodies up to the size of small asteroids¹⁷. The anisotropy of the surface temperature is established as a consequence of the hemispheric absorption of the solar radiation. Its quantitative description requires a detailed physical model of the heat conduction in the body and thermal reradiation by its surface. Interplay between the rotation and revolution periods, both observed or theoretically modelled, and the heat relaxation time-scale, derived from the measured material parameters of meteorites and observations of asteroids, results in a permanent acceleration along the orbit. The semimajor axis thus undergoes secular change at a speed and in a direction related to the body’s size, thermal properties and rotational state^{11,13,14}. Typical drift rates range from 10^{-4} to 10^{-2} AU Myr⁻¹, large enough to probably detect the Yarkovsky effect using future measurements of the near-Earth asteroid (NEA) orbits¹⁸.

As a consequence of the Yarkovsky effect, small bodies can drift in the asteroid belt for many millions of years between their initial ejection from a sizeable parent asteroid and their eventual insertion into a resonance. Such a prolonged intermediate phase implies that meteoroids are likely to be fragmented by impacts while drifting, and several generations of such break-up events may take place. Most of the original fragments delivered from a large parent body sooner or later become evolving swarms of smaller objects, all drifting at different speeds. Only a small fraction of these fragments will hit the Earth.

To describe this complex process, we have developed a realistic statistical model, including impact events along with Yarkovsky orbital drift and resonant effects. Each potential meteoroid is characterized in our simulation by three orbital (proper) elements, plus a radius, an obliquity angle and an initial time at which it starts to be irradiated by cosmic rays. We assume that all the bodies have the same thermal parameters, in particular the surface conductivity K .

Our simulations start by ejecting from a main-belt parent asteroid a swarm of fragments with a total mass equal to that of an object 500 m in diameter. The initial ejecta, as well as the fragments generated later from subsequent break-up events, have a size distribution such that the number of bodies larger than radius R is proportional to $R^{-5/2}$; their spin axes are distributed isotropically, and the orbital elements are varied according to a power-law distribution of ejection velocities¹. As most meteorites have pre-atmospheric sizes exceeding 10 cm, we always neglect in our simulations the fragments of radius less than 10 cm.

We then allow the initial swarm of about 66 million fragments to evolve by several different mechanisms that all act simultaneously. First, the semimajor axes drift according to the Yarkovsky effect, depending on the fragment’s orbital elements, size, obliquity and thermal parameters¹⁴. Second, the obliquities change randomly at average intervals $\tau_{\text{rot}} = 15 R^{1/2}$ Myr (hereafter R is in metres) and the collisional fragmentations are assumed to appear at average intervals $\tau_{\text{dis}} = 16.8 R^{1/2}$ Myr. The two lifetimes τ_{rot} and τ_{dis} are computed with a simple analytical model^{11,13,17}, using average collisional probability and impact velocity values that are appropriate for the inner part of the asteroid belt. If a body is disrupted we consider a swarm of fragments generated as described above. The cosmic-ray exposure (CRE) age of the fragments is updated by assuming that the material lying at depths greater than 1.5 m in the parent body is shielded from cosmic rays.

When a body reaches the edge of ν_6 or 3/1 resonances with the motion of the giant planets, represented by a surface in the proper-element space (and located by numerical integrations¹⁰), its further dynamics is assumed to be dominated by resonant effects and planetary close encounters (making Yarkovsky perturbations negligible). These post-resonant bodies continue to undergo shattering impacts (with $\tau_{\text{dis}} = 14.2 R^{1/2}$ Myr) and the CRE ages of the fragments are updated as above until some of their final offspring hit the Earth (we then record the radius and CRE age). For each body, this is assumed to occur after a time chosen by a Monte Carlo routine according to the distribution derived in ref. 10.

The entire simulation is stopped either after about 1 Gyr of evolution, if we aim at simulating the effects of very old break-up

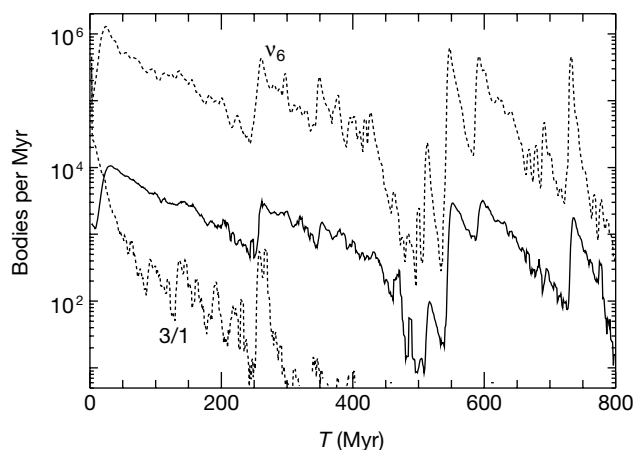


Figure 1 The expected flux of fragments from Hebe (for $K = 0.1 \text{ W m}^{-1} \text{ K}^{-1}$) versus time. Data is shown for the ν_6 and 3:1 resonances (dotted lines) and the Earth (full line). The flux is dominated by small ($R < 1 \text{ m}$) fragments, and the large fluctuations (about a factor of 100) of the resonance fluxes are a result of secondary fragmentations of relatively large bodies into swarms of smaller ones. The flux at the Earth (that is, number of impacts) mimics the behaviour of the flux to ν_6 , which is the main delivery route in this case, although it is ‘smoothed out’ by the chaotic character of the post-resonant orbits. We note that if the radius R_1 of the largest body in the initial distribution of ejecta were changed, the quantities plotted along the vertical and horizontal axes in this diagram would scale roughly proportional to $R_1^{5/2}$ and $R_1^{1/2}$, respectively.

events in the main belt; or after some given, shorter time span. In the former case, the resulting distribution of sizes and CRE ages also approximates the steady-state distribution for a population of meteoroids originating in many different events at random times over the simulation time span; the latter option appropriately simulates a specific, recent break-up event.

We have carried out nine 1-Gyr runs, starting from three different locations in the inner main belt, corresponding to ejecta swarms from the asteroids Vesta, Hebe and Flora (chosen both because they had been previously identified as probably meteorite parent bodies^{1,2,19,20} and because they are located in different regions of the orbital-element space). As the Yarkovsky effect depends on the fragments' thermal conductivity K , we have made simulations with three widely different values of this parameter: $0.0015 \text{ W m}^{-1} \text{ K}^{-1}$, as appropriate for dusty or regolith-covered surfaces²¹; $0.1 \text{ W m}^{-1} \text{ K}^{-1}$, plausible for porous or fragmented rocks^{21,22}; and $1 \text{ W m}^{-1} \text{ K}^{-1}$, a

typical value for 'bare' meteoritic rocks²². Every simulation has been repeated three times, with different choices of random seeds, and the results have been averaged to filter out any flukes.

The high efficiency of the Yarkovsky-driven transport mechanism, in synergy with the collisional cascade effects, is one of the unusual results from these simulations. About 40% to 90% of the initial ejecta mass always ends up in one of the two resonances within approximately the 1-Gyr time span (weakly depending on K and the starting location), compared to fractions of approximately 1% or less for fragments directly inserted in the resonances after the initial ejection event. Taking into account that only 1.18% and 0.23% of the ν_6 and 3/1 resonant particles hit the Earth¹⁰, our results mean that the overall Earth transfer efficiency from the main belt ranges between 0.35% and 0.86%. The corresponding absolute flux depends on the production rate of ejecta. The approximately 0.6% mean Earth delivery efficiency of ejecta from main-belt asteroids is only about a factor of 10 lower than that from NEAs²³, and as the latter are at least a factor of 100 less abundant than inner-belt bodies at sizes of the order of 1 km, kilometre-sized NEAs should not be contributing a large fraction of the meteorites.

Although the supply of meteorites to our planet is eventually about the same in our model as in the classical direct-injection scenario (because the classical model largely over-estimated the number of the Earth impactors among bodies in the resonances), from another point of view our model leads to very different conclusions. For instance, our results imply that most asteroids in the inner main belt (up to $a \approx 2.7 \text{ AU}$) must be relatively efficient sources of ejecta to the resonances, with limited dependence on the location of their orbit. This result differs from many previous studies that had indicated that only a small percentage of asteroids, especially those near the borders of major resonances, would contribute to the meteorite flux. Large asteroids, such as Vesta and Hebe, are likely to provide a higher yield due to their greater cross-section¹³, but it is possible that this is compensated by their higher escape velocities and the increasing number of asteroids at smaller and smaller sizes. A better knowledge of the asteroid/projectile size distribution and the typical ejecta speed is required to assess the relative importance of parent asteroids of different sizes^{1,13}.

Figure 1 shows that the average Earth flux from a single event decreases with respect to the early values very slowly, over times of several hundred Myr (several times longer than τ_{dis} for the largest bodies in the evolving swarm). This is consistent with the evidence that a currently abundant meteorite type, the L-chondrites, underwent a large-scale break-up event about 500 Myr ago, as indicated by their ^{39}Ar – ^{40}Ar ages²⁴. Note that the flux from any single source event fluctuates significantly, which might explain the evidence from fossil chondritic meteorites that the Earth flux was much higher some 480 Myr ago²⁵. However, any drastic fluctuations in the meteorite flux or its compositional mixture over timescales up to about 10 Myr is prevented by the chaotic nature of the post-resonant orbits, and by the fact that at any given time the Earth is hit by fragments produced by many distinct ejection events from different parent asteroids.

Our model allows us to predict the distribution of CRE ages for the Earth-hitting objects from different parent asteroids and with different values of K (Fig. 2). We note that our simulated meteorites on average have undergone two to five break-up events during the trip, suggesting that "complex" CRE histories^{15,26,27}, recording variable exposure geometries, should be commonplace. The interplay of the slow transport process and the relatively frequent fragmentations also means that in a steady-state case most CRE ages range from about 10 to 50 Myr, in agreement with the data for stony meteorites. CRE ages shorter than about 5 Myr are quite rare, and this is an important argument that favours our model over the direct injection scenario, which cannot explain why such 'young' fragments are seldom observed.

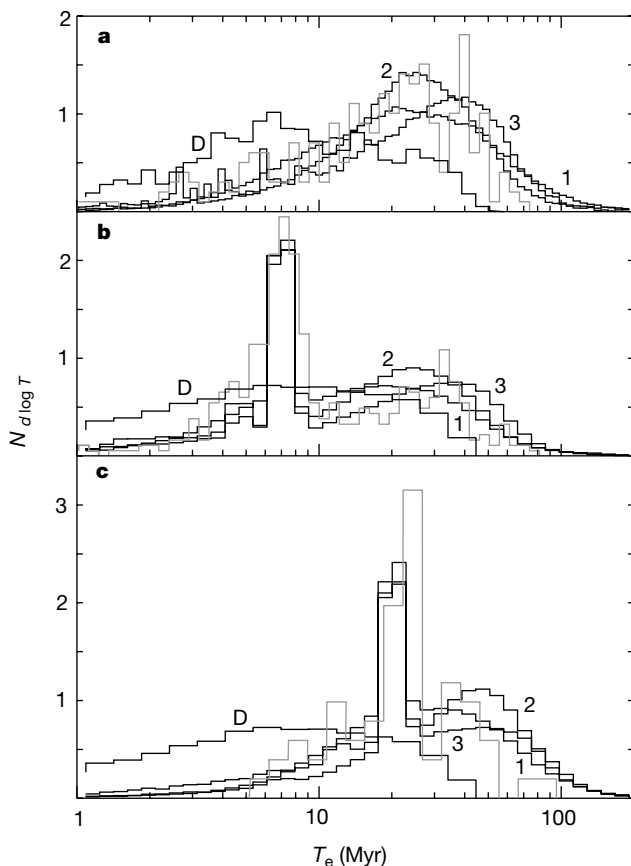


Figure 2 Comparison of the modelled and observed cosmic-ray exposure (CRE) age distributions for three different meteorite types. The data in the grey histograms is taken from refs 15 and 28. For the predictions, we show results of the direct-injection case with no Yarkovsky mobility (D histogram) and the model including Yarkovsky mobility of the meteoroids and their precursors (bold histograms). Histograms 1, 2 and 3 refer to thermal conductivity values of 0.0015 , 0.1 and $1 \text{ W m}^{-1} \text{ K}^{-1}$, respectively. As the ordinate represents an absolute quantity— $N_{d \log T_e}$: number of meteorites in a given logarithmic bin ($T_e, T_e + d \log T_e$) normalized by the total number of meteorites and the bin width $d \log T_e$ —both the data and the results of our simulations were normalized independently. **a**, Assumed ejecta from asteroid Flora whose computed CRE ages are compared with the observed distribution for 240 L-chondrites. **b**, Assumed ejecta from asteroid Hebe compared with the 444 CRE ages of H-chondrites. **c**, Assumed ejecta from asteroid Vesta, compared to the CRE age data for 64 HED (howardite–eucrite–diogenite) meteorites. In all cases, the intermediate K value appears to provide the best match to the data. We note that the direct-injection scenario would always predict many more short CRE ages than are observed, and a shortage of ages between 20 and 50 Myr. Neither of these problems is present when the Yarkovsky mobility is taken into account.

Our results indicate that the $K = 0.1 \text{ W m}^{-1} \text{ K}^{-1}$ Flora case provides a relatively good fit to the observed CRE age distribution for the most common type of stony meteorite falls, the L-chondrites (about 40% of all the falls). Flora is the largest body of a very numerous, broad 'clan' of asteroids located at $a \approx 2.2 \text{ AU}$, just out of ν_6 , so our results would suggest that this entire region of the belt is a plausible source for the L-chondrites.

It is well known that for some meteorite types the observed CRE ages show distinct clusters or peaks. In the case of the H-chondrites²⁸, about half (some 15% of all meteorite falls) have CRE ages of $7 \pm 1 \text{ Myr}$. Our model cannot reproduce these features by assuming that they are the consequence of relatively recent (but otherwise 'normal') fragment production events, because the flux from any given event lasts for hundreds of Myr, with a very slow decline as a function of time. Therefore it would be impossible to get a large fraction of the meteorites associated with a single recent event unless it were of an unusual magnitude.

To test this possibility, we ran our model again for the case of Hebe, a plausible parent body of the H-chondrites^{19,20}. We stopped the simulation after only 8 Myr and selected the bodies hitting the Earth in the last 1 Myr of this short simulation. We then superimposed these simulated discrete-event CRE ages to the steady-state distribution predicted by the model for meteorites from Hebe, with weights determined by a least-squares best fit to the data (about 0.3 and 0.7, respectively). The agreement is quite good, although there is some evidence for a second unmodelled peak at about 33 Myr.

We applied the same technique to compare the model CRE ages obtained in the case of Vesta with the data for the HED (howardite-eucrite-digenite) meteorites¹⁶ (spectral and mineralogical evidence suggests this large asteroid as the parent body^{8,29}). Here the prominent 23 Myr peak has been fitted by taking 0.2 and 0.8 weights for the discrete-event and the steady-state distributions. Although the fit of the model to the data is again good, the CRE ages longer than 50 Myr and shorter than 5 Myr are overabundant in the model histogram. This is possibly due to our simplified method of resetting the CRE clock after secondary fragmentations, which are more numerous (typically 4–5) for Vesta's ejecta because the source body is relatively far from the resonances. However, in this case the statistics are not as good, because only a small number of the CRE age data are available. □

Received 29 April 1999; accepted 9 August 2000.

- Farinella, P., Goncz, R., Froeschlé, Ch. & Froeschlé, C. The injection of asteroid fragments into resonances. *Icarus* **101**, 174–187 (1993).
- Wetherill, G. W. & Chapman, C. R. in *Meteorites and the Early Solar System* (eds Kerridge, J. F. & Matthews, M. S.) 35–67 (Univ. of Arizona Press, Tucson, 1988).
- Greenberg, R. & Chapman, C. R. Asteroids and meteorites—Parent bodies and delivered samples. *Icarus* **55**, 455–481 (1983).
- Wetherill, G. W. Asteroidal source of ordinary chondrites. *Meteoritics* **20**, 1–22 (1985).
- Wisdom, J. Meteorites may follow a chaotic route to Earth. *Nature* **315**, 731–733 (1985).
- Scholl, H. & Froeschlé, Ch. The ν_6 secular resonance region near 2 AU: A possible source of meteorites. *Astron. Astrophys.* **245**, 316–321 (1991).
- Farinella, P. et al. Asteroids falling into the Sun. *Nature* **371**, 314–317 (1994).
- Migliorini, F. et al. Vesta fragments from ν_6 and 3:1 resonances: Implications for V-type NEAs and HED meteorites. *Meteor. Planet. Sci.* **32**, 903–916 (1997).
- Gladman, B. J. et al. Dynamical lifetimes of objects injected into asteroid belt resonances. *Science* **277**, 197–201 (1997).
- Morbidelli, A. & Gladman, B. Orbital and temporal distributions of meteorites originating in the asteroid belt. *Meteor. Planet. Sci.* **33**, 999–1016 (1998).
- Farinella, P., Vokrouhlický, D. & Hartmann, W. K. Meteorite delivery via Yarkovsky orbital drift. *Icarus* **132**, 378–387 (1998).
- Hartmann, W. K. et al. Reviewing the Yarkovsky effect: New light on the delivery of stones and irons from the asteroid belt. *Meteor. Planet. Sci.* **34**, A161–A168 (1999).
- Botke, W. F., Rubincam, D. P. & Burns, J. A. Dynamical evolution of main belt meteoroids: Numerical simulations incorporating planetary perturbations and Yarkovsky thermal forces. *Icarus* **145**, 301–331 (2000).
- Vokrouhlický, D. A complete linear model for the Yarkovsky thermal force on spherical asteroid fragments. *Astron. Astrophys.* **344**, 362–366 (1999).
- Marti, K. & Graf, T. Cosmic-ray exposure history of ordinary chondrites. *Annu. Rev. Earth Planet. Sci.* **20**, 221–243 (1992).
- Welten, K. C. et al. Cosmic-ray exposure ages of diogenites and the recent collisional history of the HED parent body/bodies. *Meteor. Planet. Sci.* **32**, 891–902 (1997).

- Farinella, P. & Vokrouhlický, D. Semimajor axis mobility of asteroidal fragments. *Science* **283**, 1507–1510 (1999).
- Vokrouhlický, D., Milani, A. & Chesley, S. R. Yarkovsky effect on small near-Earth asteroids: Mathematical formulation and examples. *Icarus* (in the press).
- Migliorini, F. et al. Surface properties of (6) Hebe: A possible parent body of ordinary chondrites. *Icarus* **128**, 104–113 (1997).
- Gaffey, M. J. & Gilbert, S. L. Asteroid 6 Hebe: The probable parent body of the H-type ordinary chondrites and the IIE iron meteorites. *Meteor. Planet. Sci.* **33**, 1281–1295 (1998).
- Presley, M. A. & Christensen, P. R. Thermal conductivity measurements of particulate materials. *J. Geophys. Res.* **102**, 6535–6550 (1997).
- Yomogida, K. & Matsui, T. Physical properties of ordinary chondrites. *J. Geophys. Res.* **88**, 9513–9533 (1983).
- Gladman, B., Michel, P. & Froeschlé, Ch. The near-Earth object population. *Icarus* **146**, 176–189 (2000).
- Haack, H., Farinella, P., Scott, E. R. D. & Keil, K. Meteoritic, asteroidal and theoretical constraints on the 500 Ma disruption of the L chondrite parent body. *Icarus* **119**, 182–191 (1996).
- Schmitz, B., Peucker-Ehrenbrink, B., Lindström, M. & Tassinari, M. Accretion rates of meteorites and cosmic dust in the early Ordovician. *Science* **278**, 88–90 (1997).
- Wetherill, G. W. Multiple cosmic-ray exposure ages of meteorites. *Meteoritics* **15**, 386–387 (1980).
- Herzog, G. F. et al. Complex exposure histories for meteorites with "short" exposure ages. *Meteor. Planet. Sci.* **32**, 413–422 (1997).
- Graf, T. & Marti, K. Collisional history of H chondrites. *J. Geophys. Res.* **100**, 21247–21263 (1995).
- Binzel, R. P. & Xu, S. Chips off of asteroid 4 Vesta—Evidence for the parent body of basaltic achondrite meteorites. *Science* **260**, 186–191 (1993).

Acknowledgements

This Letter is dedicated to the memory of P.F., who passed away on 25 March 2000. Our understanding of the role of the Yarkovsky effect in the history of the Solar System is only one of many contributions by P.F. to modern planetology. We thank J. A. Burns and C. R. Chapman for comments on the manuscript.

Correspondence and requests for materials should be addressed to D.V.
(e-mail: vokrouhl@mbox.cesnet.cz).

Trapping and emission of photons by a single defect in a photonic bandgap structure

Susumu Noda, Alongkarn Chutinan & Masahiro Imada

Department of Electronic Science and Engineering, Kyoto University,
Kyoto 606-8501, Japan

By introducing artificial defects and/or light-emitters into photonic bandgap structures^{1,2}, it should be possible to manipulate photons. For example, it has been predicted² that strong localization (or trapping) of photons should occur in structures with single defects, and that the propagation^{3,4} of photons should be controllable using arrays of defects. But there has been little experimental progress in this regard, with the exception of a laser⁵ based on a single-defect photonic crystal. Here we demonstrate photon trapping by a single defect that has been created artificially inside a two-dimensional photonic bandgap structure. Photons propagating through a linear waveguide are trapped by the defect, which then emits them to free space. We envisage that this phenomenon may be used in ultra-small optical devices whose function is to selectively drop (or add) photons with various energies from (or to) optical communication traffic. More generally, our work should facilitate the development of all-optical circuits incorporating photonic bandgap waveguides and resonators.

The photonic bandgap (PBG) structure considered in this work is a two-dimensional PBG slab with a triangular lattice structure as shown in Fig. 1a. The structure makes use of the effect of a two-dimensional PBG to confine the light in the in-plane direction for transverse-electric (TE)-like mode, and large refractive index

contrast to confine the light in the vertical direction⁵. The radius of each hole and the thickness of the slab are chosen to be $0.29a$ and $0.6a$, respectively, where a is the lattice constant of the two-dimensional structure. In Fig. 1a, line-shaped defects are introduced to form a straight waveguide. We calculated the transmission property of the waveguide with the three-dimensional finite-difference time-domain (FDTD) method⁶ with the boundary conditions of ref. 7 and confirmed that lossless transmission can be obtained in a wide range of frequency at approximately $(0.27-0.28)(c/a)$, where c is the velocity of light in the vacuum⁸. Here, we introduce a single defect in the vicinity of the waveguide as shown in Fig. 1b, where the radius of the defect and the distance between the defect and waveguide are $0.56a$ and $(3\sqrt{3}/2)a$, respectively. In this case, the defect acts as an optical resonator whose resonant frequency f_i and quality factor Q are calculated to be $0.2729(c/a)$ and about 500, respectively. When $a = 0.42 \mu\text{m}$, the normalized frequency f_i corresponds to the wavelength of $1.539 \mu\text{m}$. The Q value is determined by the following two quality factors: Q_{in} for in-plane direction and Q_{v} for vertical direction. Q_{in} is primarily determined by the distance between the defect and the waveguide, and it becomes larger with an increasing distance. On the other hand, Q_{v} is determined by the effective refractive index contrast between the defect and the cladding air, and it can be tuned by changing, for example, the thickness of the slab⁹. In the present geometry (Fig. 1b), Q_{in} and Q_{v} are almost equal.

The flux of photons which are trapped from the waveguide and emitted to free space by the single defect was calculated by the three-dimensional FDTD method, and the result is shown in Fig. 1c. It is clearly seen that the photons are indeed emitted to free space via the

single defect for a resonant frequency of $f_i = 0.2729(c/a)$ ($1.539 \mu\text{m}$ for $a = 0.42 \mu\text{m}$). From detailed theoretical consideration, we have found that half of the photon flux which flows through the waveguide can be trapped and emitted to free space, when the values of Q_{in} and Q_{v} are equal. When a large mismatch between the values of Q_{in} and Q_{v} is present, the emitted photon flux becomes smaller. These phenomena can be understood using the analogy of the impedance matching condition in an electronic circuit: the energy transfer from the circuit to the load becomes maximal when the output impedance of the electronic circuit is equal to the impedance of the load.

Having shown that trapping and emission of photons by a single defect are indeed possible, we next considered introducing an additional defect, as shown in Fig. 2a. The radius of the additional defect is assumed to be $0.58a$, which is about 3.5% larger than that of the defect described above, and its resonant frequency f_j is $0.2769(c/a)$ (that is, $1.517 \mu\text{m}$ for $a = 0.42 \mu\text{m}$). As can be seen in our calculated results of Fig. 2b, photons with different frequencies f_i and f_j are trapped and emitted to free space by the corresponding defects. The result clearly shows that the drop (or add) function of photons can be realized by introducing multiple defects with different resonant frequencies at the vicinity of the waveguide.

In light of these promising results, we then fabricated a device (see scanning electron micrograph (SEM) in Fig. 3). The device was fabricated as follows. First, an InGaAsP layer with thickness of $0.25 \mu\text{m}$ and electronic bandgap wavelength of $1.1 \mu\text{m}$ was grown on InP substrate by metal-organic vapour phase epitaxy (MOVPE). A two-dimensional triangular lattice structure with a lattice constant

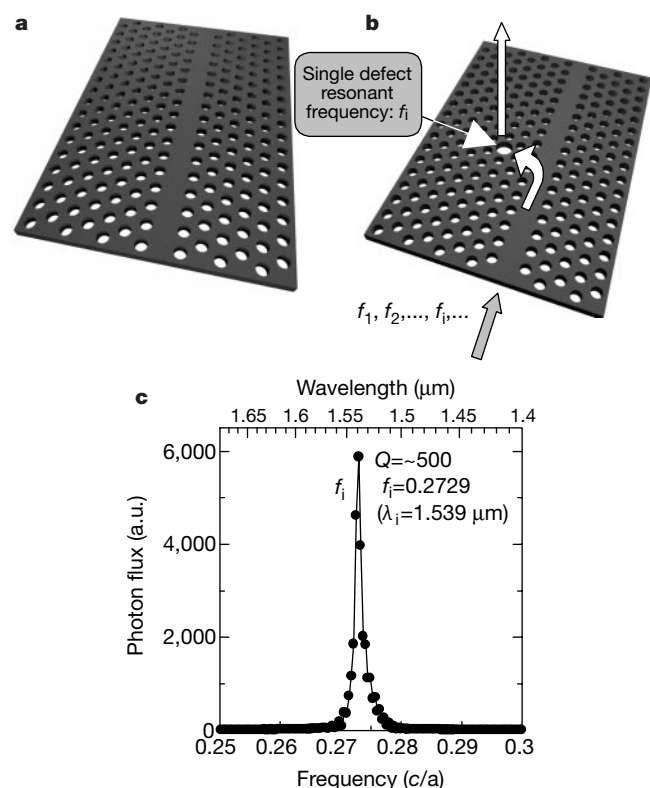


Figure 1 Trapping and emission of photons by a single defect in a photonic bandgap (PBG). **a**, Two-dimensional triangular lattice slab used as the base PBG structure. The radius of each hole is $0.29a$ and the thickness of slab is $0.6a$, where a is the lattice constant. Line-shaped defects are introduced to form a straight waveguide. From the theoretical calculation, the lossless transmission can be expected in a frequency region about $(0.27-0.28)(c/a)$. **b**, A single defect formed at the vicinity of the waveguide; radius $r_i = 0.56a$ and resonant frequency $f_i = 0.2729(c/a)$ (that is, $1.539 \mu\text{m}$ for $a = 0.42 \mu\text{m}$). **c**, Calculated photon flux trapped from the waveguide and emitted to free space by the single defect. The top axis shows the wavelength for $a = 0.42 \mu\text{m}$.

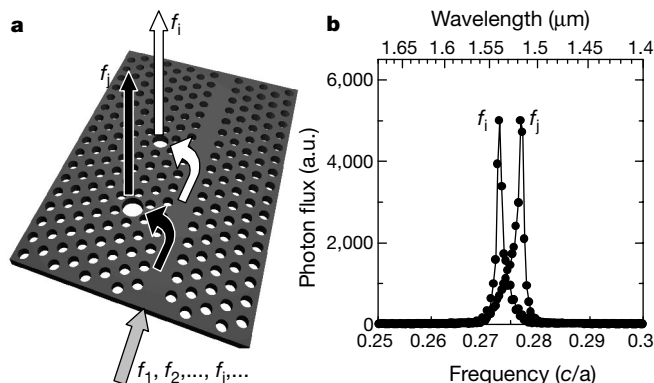


Figure 2 Trapping and emission of photons by two isolated defects. **a**, A new defect with radius $r_j = 0.58a$ and resonant frequency $f_j = 0.2769(c/a)$ (that is, $1.517 \mu\text{m}$ for $a = 0.42 \mu\text{m}$) is added to the defect described in Fig. 1b. **b**, Calculated photon flux emitted by two defects. The top axis shows the wavelength for $a = 0.42 \mu\text{m}$. Photons with different frequencies of f_i and f_j are trapped and emitted to free space by the corresponding defects. In the calculation, the mutual interaction between two defects is not considered.

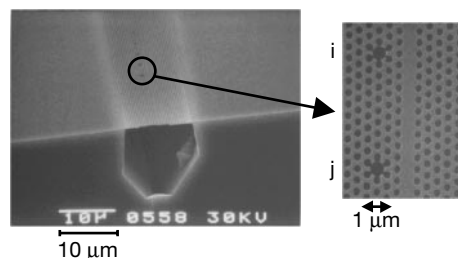


Figure 3 Scanning electron microscopy picture of the fabricated sample. The InGaAsP two-dimensional slab structure surrounded by air in the vertical direction is fabricated, on which a straight waveguide and isolated defects labelled i and j are formed.

of $a = 0.42 \mu\text{m}$, together with line-shaped defects (that is, the waveguide) and isolated defects with different diameters, were drawn on a resist mask coated on the InGaAsP layer by electron beam lithography. The PBG wavelength of the two-dimensional slab is designed to be approximately in the $1.55\text{-}\mu\text{m}$ region, which is especially important for fibre-optic communication. The resist pattern was then transferred to the InGaAsP layer by a reactive ion etching technique, and the InP substrate under the patterned InGaAsP layer was selectively etched-off to form the slab structure. Figure 3 shows the successful fabrication of the InGaAsP two-dimensional slab structure with isolated defects and the waveguide. Here, we concentrate on the two defects labelled i and j, which are shown in the inset to Fig. 3. The difference in the radii of the defects i and j is 3–4% ($j > i$).

The experimental results for the trapping and emission of photons by two defects i and j are shown in Fig. 4. Figure 4a shows the top view of the device. The photons were injected to the waveguide by an optical fibre with a lens from the right edge of the slab. The light source was a semiconductor laser whose oscillation wavelength can be tuned around the $1.55\text{-}\mu\text{m}$ region, which corresponds to the normalized frequency region of about $0.2710(c/a)$. The trapping and emission phenomena were observed by infrared camera, and the results are shown in Fig. 4b. When the frequency of the light source is tuned at $f = 0.2718(c/a)$ (that is, $1.545 \mu\text{m}$), defect j emits light strongly, and defect i does not. On the

other hand, when the frequency is tuned at $f = 0.2682(c/a)$ (that is, $1.566 \mu\text{m}$), emission from defect j disappears but strong emission from defect i appears. The difference in the measured resonant frequencies ($\Delta f = 0.0036(c/a)$) of the defects i and j are very close to that of the theoretical value discussed above ($\Delta f = 0.004(c/a)$).

Next, we characterized the localized photon state for each defect. For this purpose, we prepared another sample with only a single defect, corresponding to defect i of Fig. 4. In our present measurement system, it is difficult to distinguish the spectra spatially from the individual defects shown in Fig. 4 because they are very close to each other ($\sim 4 \mu\text{m}$). Thus, we prepared another sample to measure the photon state of a pure single defect. The emitted power from the defect as a function of wavelength is plotted in Fig. 5. A peak with a centre frequency of $0.2686(c/a)$ (that is, $1.563 \mu\text{m}$) and a linewidth of $0.0007(c/a)$ (that is, 4.1 nm) can be clearly seen. We consider the non-zero background to be due to the light scattered at the entrance (or exit) edge of the waveguide (Fig. 4b). We can estimate the Q value of the defect to be about 400 from Fig. 5, very close to the theoretical estimation described above. Another important characteristic is the coupling efficiency from the in-plane to the vertical direction by the defect. Although it is difficult to obtain an exact value for the coupling efficiency, we estimate it to be roughly a few tens of per cent, on the basis of the fact that the intensity of the light scattered at the exit edge of the waveguide became weaker by at least a few tens of per cent when strong emission was observed at the defect. This indicates that very high coupling efficiency can be expected using our method. As described above, the coupling efficiency of 50% at maximum is expected when the Q_{in} and Q_{v} are equal. These results clearly show that the trapping and emission of photons can indeed be achieved by introducing single defects at the vicinity of the waveguide in a PBG structure.

Our results also provide a new way to develop ultra-small optical devices. Because the size of the defect is of the order of one wavelength and the photons are emitted in a direction normal to the surface, the structure can be applied to a very compact surface-emitting-type wavelength-add-drop (or wavelength-monitoring) device, which is important for wavelength division multiplex optical communication. The Q value of the defect can be controlled by the geometry of the structure as described above, and thus the frequency spacing for dropping (or adding) can be controlled arbitrarily. Also, the output power from the defect can be controlled by adjusting the relationship between the values of Q_{in} and Q_{v} . Other possible applications using strong localization of photons at the defect include enhancement of nonlinear optical phenomena and trapping of nanoparticles. □

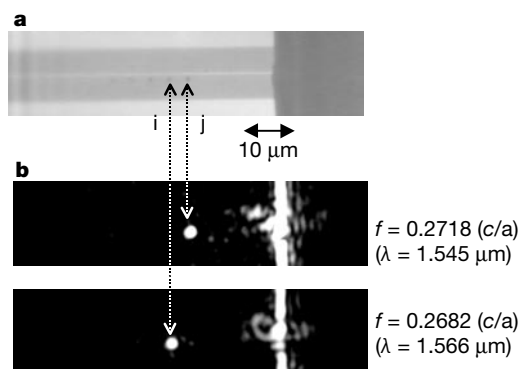


Figure 4 Experimental results of trapping and emission of photons by defects. **a**, Top view of the sample observed by optical microscope, where the waveguide and defects i and j are seen. The light was injected to the waveguide from the right edge of the slab. **b**, Experimental results observed by infrared camera. When the input light frequency is tuned at $f = 0.2718(c/a)$ (that is, $1.545 \mu\text{m}$), defect j emits the light strongly but defect i does not emit. On the other hand, when the frequency is tuned at $f = 0.2682(c/a)$ (that is, $1.566 \mu\text{m}$), the reverse phenomenon occurs.

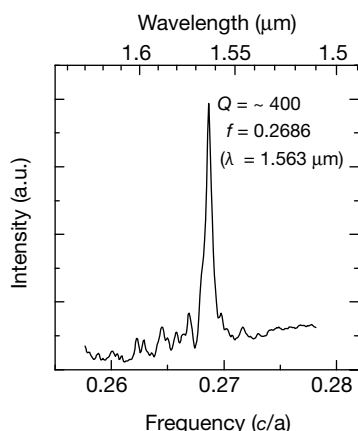


Figure 5 Emitted photon flux from a single defect as a function of normalized frequency (or wavelength). The total Q of the localized photon state is estimated to be ~ 400 , which corresponds well to the theoretical calculation.

Received 10 April; accepted 17 July 2000.

- Yablonovitch, E. Inhibited spontaneous emission in solid-state physics and electronics. *Phys. Rev. Lett.* **58**, 2059–2062 (1987).
- John, S. Strong localization of photons in certain disordered dielectric superlattices. *Phys. Rev. Lett.* **58**, 2486–2489 (1987).
- Joannopoulos, J. D., Meade, R. D. & Winn, J. N. *Photonic Crystals* (Princeton, New York, 1995).
- Fan, S., Villeneuve, P. R. & Joannopoulos, J. D. Channel drop tunneling through localized states. *Phys. Rev. Lett.* **80**, 960–963 (1998).
- Painter, O. J. et al. Two-dimensional photonic band-gap defect mode laser. *Science* **284**, 1819–1821 (1999).
- Yee, K. S. Numerical solution of boundary value problems involving Maxwell's equations in isotropic media. *IEEE Trans. Antennas Propag.* **14**, 302–307 (1966).
- Mur, G. Absorbing boundary conditions for the finite-difference approximation of the time-domain electromagnetic-field equations. *IEEE Trans. Electromagn. Compat.* **23**, 377–382 (1981).
- Chutinan, A. & Noda, S. Waveguides and waveguide bends in two-dimensional photonic crystal slabs. *Phys. Rev. B* **62**, 4488–4492 (2000).
- Painter, O. J., Vuckovic, J. & Scherer, A. Defect modes of a two-dimensional photonic crystal in an optically thin dielectric slab. *J. Opt. Soc. Am. B* **16**, 275–280 (1999).

Acknowledgements

This work is partly supported by a grant-in-aid for scientific research of priority areas from the Ministry of Education, Science and Culture of Japan.

Correspondence and requests for materials should be addressed to S. N. (e-mail: snoda@kuee.kyoto-u.ac.jp).

Colloidal ordering from phase separation in a liquid-crystalline continuous phase

Jean-Christophe Loudet, Philippe Barois & Philippe Poulin

Centre de Recherche Paul Pascal / CNRS, Avenue Schweitzer, 33600 Pessac, France

Some binary mixtures exist as a single phase at high temperatures and as two phases at lower temperatures; rapid cooling therefore induces phase separation that proceeds through the initial formation of small particles and subsequent growth and coarsening¹. In solid and liquid media, this process leads to growing particles with a range of sizes, which eventually separate to form a macroscopically distinct phase. Such behaviour is of particular interest in systems composed of an isotropic fluid and a liquid crystal², where the random distribution of liquid-crystal droplets in an isotropic polymer matrix may give rise to interesting electro-optical properties. Here we report that a binary mixture consisting of an isotropic fluid and a liquid crystal forming the continuous phase does not fully separate into two phases, but self-organizes into highly ordered arrays of monodisperse colloidal droplet chains. We find that the size and spatial organization of the droplets are controlled by the orientational elasticity of the liquid-crystal phase and the defects caused by droplets exceeding a critical size. We expect that our approach to forming monodisperse, spatially ordered droplets in liquid crystals will allow the controlled design of ordered composites that may have useful rheological and optical properties.

The system studied here is a mixture of poly(dimethylsiloxane-co-methylphenylsiloxane) ('silicone oil', Aldrich) and nematic liquid crystal ('E7', a mixture of cyano-biphenyl and terphenyl molecules, Merck); the nematic phase of E7 is stable over a large temperature range around room temperature. A part of the phase diagram of the mixture is shown in Fig. 1. The system forms an isotropic (I) phase at high temperature, and a biphasic equilibrium between an isotropic and a nematic phase (I+N) at low temperature. A nematic domain (N) is found at intermediate temperatures and low concentrations of silicone oil. The present mixture exhibits classical features usually observed in mixtures of nematic liquid crystal and isotropic fluids^{3,4}.

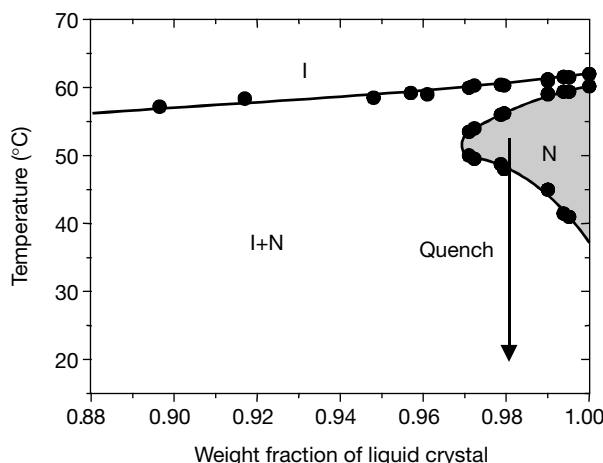


Figure 1 Phase diagram of the mixture of liquid crystal and silicone oil. The figure shows the isotropic phase (I), equilibrium between an isotropic and a nematic phase (N+I), and the nematic phase (N). The systems are quenched from the nematic domain to 20 °C, as indicated by the arrow.

We have performed temperature quenches of this mixture from the nematic domain to the biphasic domain of the phase diagram—quenches from the isotropic phase would have led to the formation of numerous and uncontrolled topological defects⁵. The mixture under investigation was contained between two glass slides, separated by 20- μm -thick spacers. Before the mixture was added, the glass was coated with polyvinylalcohol and rubbed: this caused the nematic continuous phase to become homogeneously aligned parallel to the slides⁶ after being kept at 55 °C for a few minutes. Once aligned, the system was quenched into the biphasic domain at 20 °C by placing the sample onto a temperature-controlled stage of an optical microscope.

Figure 2 shows the system at different times after the temperature quench. As in classical solid or fluid mixtures, small isotropic particles nucleate in the continuous liquid-crystalline phase soon after the quench. These particles, which are mostly composed of silicone oil, diffuse and grow through coalescence for a few tens of seconds. In contrast to previously known phase separation mechanisms¹, coalescence suddenly stops as the isotropic droplets reach a critical size that is about a few micrometres. Then the droplets attract each other at long range, forming long linear chains that are parallel to the alignment direction of the liquid-crystalline phase. The chains then grow through 'tip to head' aggregation, and self-organize with respect to each other; ultimately, they form highly ordered arrays of macroscopic chains (several millimetres long) made of monodisperse droplets which do not coalesce. Examples of resultant systems are shown in Fig. 3a and b. Furthermore, Fig. 3c shows the versatility of the present process: controlled patterns may be made, with domains aligned in different directions.

Most of these features can be explained by considering the specific interactions that arise from the orientational elasticity⁷ of the liquid-crystal phase. A spherical particle suspended in liquid crystal tends to distort the long-range ordering of the nematic phase⁷. Assuming a single elastic constant K , for splay, twist and bend deformations⁶, the elastic cost of the distortions around a single particle should be of the order of Ka , where a is the particle size. These distortions, which mediate elastic interactions between the particles, depend critically on the boundary conditions at the surface of the particles. These boundary conditions are spontaneously set by molecular interactions between the liquid-crystal

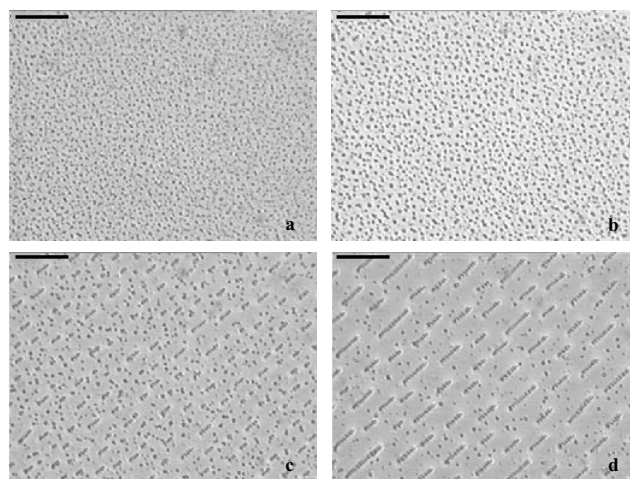


Figure 2 The phase separation process in a liquid-crystalline fluid. System composition: liquid crystal, 98 wt%; silicone oil, 2 wt%. **a**, Some droplets form after the quench (picture: 20 s after the quench). **b**, The droplets diffuse randomly and coalesce in the initial stages of the separation (picture: 42 s after the quench). **c**, Coalescence stops after a critical size is reached. The droplets form small linear aggregates oriented in the alignment direction of the liquid crystal (picture: 55 s after the quench). **d**, The chains grow as time elapses (picture: 120 s after the quench). Scale bars, 60 μm .

molecules and the silicone oil interface. The preferential orientation of the liquid-crystal molecules here is normal to the oil interface. The characteristic energy specifying a deviation from this preferential anchoring is given by wa^2 , where w is a surface energy term. For small particles, this contribution is weaker than Ka , the bulk elastic energy^{8–10}. As a consequence, small particles do not induce strong distortions of the liquid-crystalline medium (Fig. 4) and the liquid crystal adopts different orientations at the surface of the droplet. As experimentally observed in the early stages of the phase separation, such droplets can freely diffuse and coalesce when they collide with each other.

Different behaviour is expected at longer times, when the droplets become larger. For particles typically larger than K/w , the surface energy becomes dominant and the liquid crystal is expected to preserve a normal orientation at the surface of the droplet. Under these conditions, different distortions are expected⁷. As shown in Fig. 4, the far-field alignment of the liquid crystal can be satisfied by the formation of a companion hyperbolic defect located near the particle^{7,9–12}. By analogy with electrostatics^{11,13}, the companion-defect pairs are expected to behave like electrostatic dipoles at long range. This explains the formation of long chains in which the droplet-defect pairs are pointing in the same direction. The resultant chains are oriented along the alignment direction of the liquid crystal. The presence of a topological defect between two neighbouring droplets induces a short-range repulsion that prevents the droplets from approaching too closely; this inhibits their coalescence. Assuming that w is of the order of $5 \times 10^{-6} \text{ J m}^{-2}$ (typical surface anchoring energy for cyanobiphenyl molecules and silicone material¹⁴) and that K is of the order of 10 pN (ref. 6), we expect the size of the droplets to be of the order of 2 μm , a value very close to that experimentally observed. This rough estimate provides an additional support to the described mechanism. Finally, we stress that the crossover between the distinct behaviours of large and small droplets is sharp enough to lead to remarkably uniform dispersions. Varying the constituents of the mixture, or adding surface-active

compounds, can result in different anchoring energies and lead to differently sized particles. Using this approach in some preliminary experiments, we have obtained large monodisperse particles with diameter greater than 10 μm . Such particles were produced by using a different silicone oil (silicone oil DC 710, Fluka), for which w is weaker.

Dipolar elastic interactions between colloidal particles in liquid crystals are reminiscent of the behaviour of electro-rheological fluids^{7,15}. By analysing series of digitized pictures, we found that the average length of the chains grows as $t^{0.55 \pm 0.10}$, where t is time. We choose $t = 0$ to be when the initial coalescence stops and uniformly sized droplets start to aggregate. The growth rate in nematic liquid crystals is in good agreement with that found in electro-rheological fluids^{15–18}. However, a significant difference is observed at long times. In electro-rheological fluids, long chains aggregate laterally and assemble into larger columns^{18,19}. In nematic fluids, the chains do not aggregate in this way, but rather form highly ordered arrays that span the whole surface area of the sample. This behaviour is the manifestation of a long-range elastic repulsion between the chains, in sharp contrast with electrostatic systems.

By analogy with known materials—such as electro-rheological fluids¹⁵, gel networks of colloids in liquid crystals²⁰ and monodisperse ferrofluid emulsions²¹—the present systems may exhibit unique rheological and optical properties. In addition, the physical picture described in this work could help the adjustment of different parameters in order to get a specific array of monodisperse particles, and control the resultant physical properties. The alignment direction of the droplet chains, imposed by the liquid crystal, could be controlled either by mechanical or optical surface treatments or simply by external fields. We have performed preliminary experiments that confirm it is possible to control the orientation of the

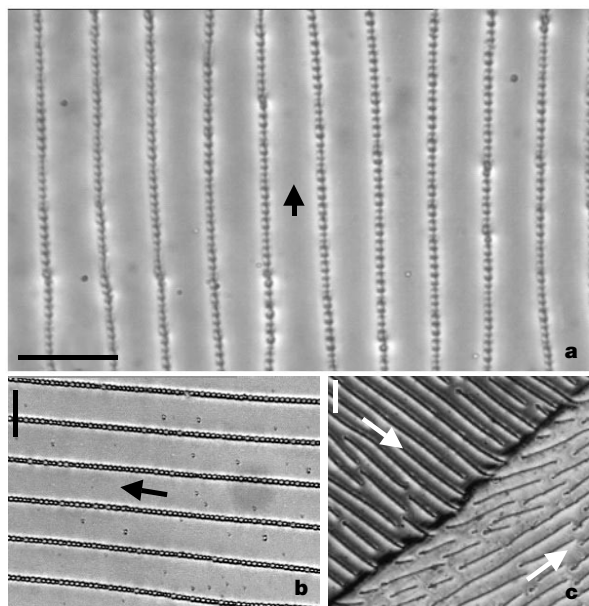


Figure 3 Very long chains are obtained at long times. They form highly ordered arrays made of monodisperse droplets. **a**, System composition: liquid crystal, 98.0 wt%; silicone oil, 2.0 wt%. Black arrow, direction of polymer rubbing. Scale bar, 50 μm . **b**, System composition: liquid crystal, 98.4 wt%; silicone oil, 1.6 wt%. Black arrow, direction of polymer rubbing. Scale bar, 35 μm . **c**, Chains aligned in different directions (picture of system between crossed polarizers). They are obtained by rubbing different regions of the substrate in different directions. System composition: liquid crystal, 98.4 wt%; silicone oil, 1.6 wt%. White arrows, directions of polymer rubbing. Scale bar, 60 μm .

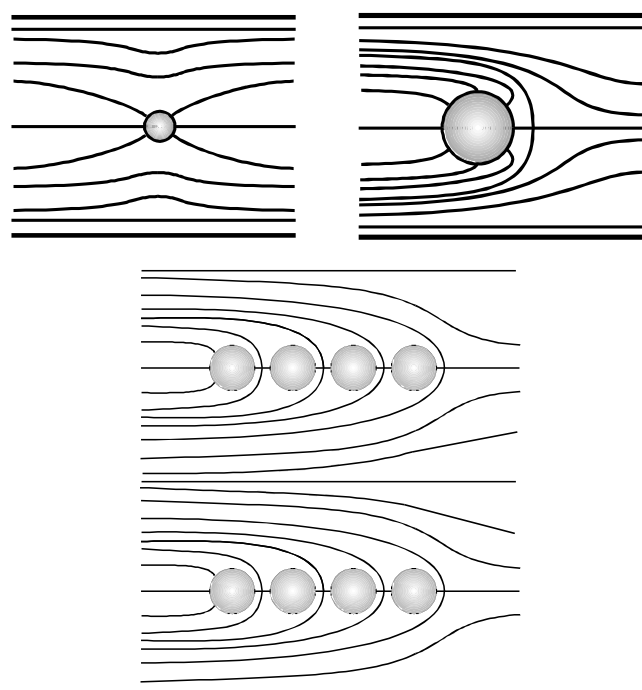


Figure 4 Schematic diagram of the distortions around particles in nematic liquid crystal. The black lines represent the axis of preferential orientation of the liquid-crystal molecules. For small particles (top left) the liquid crystal is slightly distorted, and its alignment at the surface of the particle may adopt different orientations. For large particles (top right), the alignment is normal to the particle surface. This induces the formation of a companion defect located near the particle. The global structure has dipolar symmetry. The bottom drawing represents the distortions for an assembly of dipolar droplets that form parallel chains. The defect between neighbouring droplets inhibits the coalescence.

chains by applying a weak electrical field. Finally, phase separation in liquid-crystalline fluids could be used to make monodisperse liquid or even solid particles (using cross-linkable fluids or polymers with low glass transition temperatures). □

Received 10 April; accepted 7 August 2000.

1. Gunton, J. M., San Miguel, M. & Sahni, P. S. in *Phase Transition and Critical Phenomena* Vol. 8 (ed. Lebowitz, J. L.) 267 (Academic, London, 1983).
2. Drzaic, P. *Liquid Crystal Dispersions* (Series on Liquid Crystals Vol. 1, World Scientific, Singapore, 1995).
3. Benmouna, F., Bedjaoui, L., Maschke, U., Coqueret, X. & Benmouna, M. On the phase behavior of blends of polymers and nematic liquid crystals. *Macromol. Theory Simul.* **7**, 599–611 (1998).
4. Shen, C. & Kyu, T. Spinodals in a polymer dispersed liquid crystal. *J. Chem. Phys.* **102**, 556–562 (1995).
5. Yurke, B., Pargellis, A. N., Chuang, I. & Turok, N. Coarsening dynamics in nematic liquid crystals. *Physica B* **178**, 56–72 (1992).
6. De Gennes, P. G. & Prost, J. *The Physics of Liquid Crystals* 2nd edn (Clarendon, Oxford, 1993).
7. Poulin, P., Stark, H., Lubensky, T. C. & Weitz, D. A. Novel colloidal interactions in anisotropic fluids. *Science* **275**, 1770–1773 (1997).
8. Mondain-Monval, O., Frances, N. & Poulin, P. Weak surface energy in nematic dispersions: Saturn ring defects and quadrupolar interactions. *Eur. Phys. J. B* **12**, 167–170 (1999).
9. Ruhwandl, R. W. & Terentjev, E. M. Monte Carlo simulation of topological defects in the nematic liquid crystal matrix around a spherical colloid particle. *Phys. Rev. E* **56**, 5561–5565 (1997).
10. Stark, H. Director field configurations around a spherical particle in a nematic liquid crystal. *Eur. Phys. J. B* **10**, 311–321 (1999).
11. Lubensky, T. C., Petey, D. & Currier, N. Topological defects and interactions in nematic emulsions. *Phys. Rev. E* **57**, 610–625 (1998).
12. Poulin, P. & Weitz, D. A. Inverted and multiple nematic emulsions. *Phys. Rev. E* **57**, 626–637 (1998).
13. Poulin, P., Cabuil, V. & Weitz, D. A. Direct measurement of colloidal forces in an anisotropic solvent. *Phys. Rev. Lett.* **79**, 4862–4865 (1997).
14. Lavrentovich, O. D., Nazarenko, V. G., Sergan, V. V. & Durand, G. Dielectric quenching of the electric polar surface instability in a nematic liquid crystal. *Phys. Rev. A* **45**, R6969–R6972 (1992).
15. Halsey, T. C. & Toor, W. Structure of electrorheological fluids. *Phys. Rev. Lett.* **65**, 2820–2823 (1990).
16. Miyazima, S., Meakin, P. & Family, F. Aggregation of oriented anisotropic particles. *Phys. Rev. A* **36**, 1421–1427 (1987).
17. See, H. & Doi, M. Aggregation kinetics in electrorheological fluids. *J. Phys. Soc. Jpn* **60**, 2778–2782 (1991).
18. Carmen Miguel, M. & Pastor-Satorras, R. Kinetic growth of field-oriented chains in dipolar colloidal solutions. *Phys. Rev. E* **59**, 826–834 (1999).
19. Halsey, T. C. & Toor, W. Fluctuation-induced couplings between defect lines or particle chains. *J. Statist. Phys.* **61**, 1257–1281 (1990).
20. Meeker, S. P., Poon, W. C. K., Crain, J. & Terentjev, E. M. Colloid liquid crystal composites: An unusual soft solid. *Phys. Rev. E* **61**, R6083–R6086 (2000).
21. Calderon, F. L., Stora, T., Monval, O. M., Poulin, P. & Bibette, J. Direct measurement of colloidal forces. *Phys. Rev. Lett.* **72**, 2959–2962 (1994).

Correspondence and requests for materials should be addressed to P.P. (e-mail: poulin@crpp.u-bordeaux.fr).

Patterning of polymer-supported metal films by microcutting

Natalie Stutzmann*, Theo A. Tervoort*, Kees Bastiaansen*† & Paul Smith*

* Department of Materials, ETH Zentrum, UNO-C15, CH-8092 Zürich, Switzerland

† Eindhoven University of Technology, Eindhoven Polymer Laboratories, PO Box 513, 5600 MB Eindhoven, The Netherlands

The ability to micropattern materials is of great importance for manufacturing advanced electronic, optical and mechanical devices ranging from displays to biosensors^{1–6}. For this purpose a variety of methods have been developed, including X-ray, electron-beam and photo-lithography^{7,8}, microcontact printing⁹, embossing^{10,11}, micromoulding^{8,12} and cold welding¹³. But these techniques are often of restricted applicability, involve a multitude of elaborate and cumbersome processing steps, or require aggressive chemistry. Here we describe a simple and versatile way

to create well resolved metallic structures on polymer substrates, which is based on solid-state embossing of metal-coated polymer films. Ductility of both the metal layer and the polymer substrate permits the metal to be cut into surprisingly regular, micrometre-size structures. We illustrate the method by preparing patterned electrically conducting structures, highly efficient infrared polarizers and polarization-dependent colour filters.

The procedure for generating micro-engineered metal/polymer structures is outlined in Fig. 1. Polymer films coated with a thin (5–150 nm) metal layer were embossed with a silicon 'master' in their solid state: below the melting temperature T_m but above the glass transition T_g of semi-crystalline polymer substrates, or at around T_g for amorphous polymer substrates. We used films of poly(tetrafluoroethylene-co-hexafluoropropylene) (FEP; with $T_m = 260^\circ\text{C}$ and $T_g = 80^\circ\text{C}$) coated with gold or an 80:20 per cent by weight gold–palladium alloy. We embossed them by applying a nominal pressure on the master of 850 g mm^{-2} for 5 min at 200°C , after which the bilayer films were cooled to room temperature and the load was removed. A summary of various structures obtained is presented in Fig. 2. The environmental scanning electron micrograph of a microcut 50-nm gold/FEP bilayer (Fig. 2a) shows the potential for high accuracy of the process. However, cutting of relatively thin gold layers (<20 nm) on the FEP substrate caused breaking-up of the metal, because the film thickness approached the size of its grains (5–25 nm; Fig. 2b). This was less pronounced for the finer-grained gold–palladium alloy^{14,15} (Fig. 2c). We observed that gold–palladium layers as thin as 8 nm could be microcut without failure.

It seems that reduced metal-grain size decreases the minimum thickness of the metal layer required for failure-free microcutting. It was found, though, that gold–palladium layers of about 50 nm had a greater tendency to crack than those of gold (Fig. 2d), which is also attributable to the reduced grain size of the alloy and accompanied

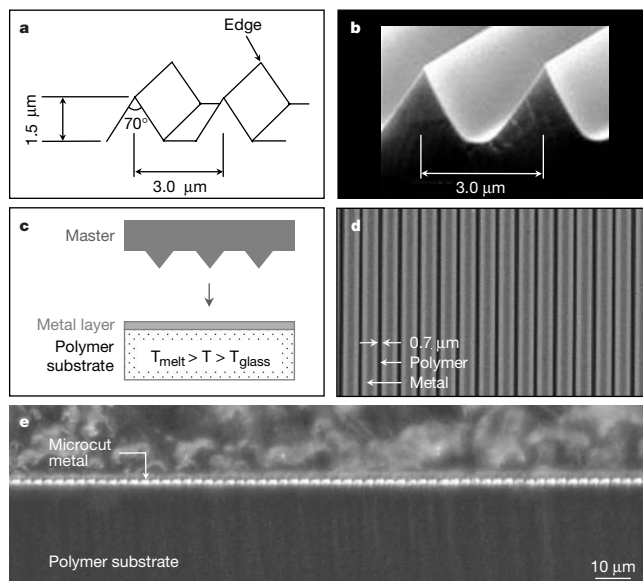


Figure 1 Diagram of the microcutting procedure of metal-coated polymer substrates. A silicon master of zig-zag groove geometry (**a**, **b**) was employed for embossing the metal/polymer bilayer, while the polymer is in its solid state (**c**), yielding regular metal microstructures on the polymer substrate. An environmental scanning electron (ESE) micrograph of a representative micro-engineered metal/polymer surface obtained with this procedure is shown in **d**, **e**. Optical micrograph of a cross-section ($\sim 10\text{ }\mu\text{m}$ thick) of a microcut metal/polymer bilayer film taken in transmitted light with crossed polarizers. Photographs (**d**, **e**) illustrate the highly regular and complete piercing of the metal layer.

reduced ductility^{16,17}. The optimum thickness for microcutting gold–palladium layers on FEP appeared to be approximately 25 nm, and for gold about 50 nm. Unexpectedly, when the thickness of the gold layer exceeded 50 nm, it was uniformly cut in every second groove over extended areas (Fig. 2e, f). In other experiments, films of high-density polyethylene (HDPE) were used as support. Gold layers evaporated onto this polymer showed a smoother morphology than those on FEP; and 20-nm-thin gold layers were also patterned without failure. The ‘doubling’ effect was noted with HDPE samples, again for 100-nm gold layers. Using oxygen-plasma-treated HDPE substrates, however, virtually eliminated this phenomenon. As polymer surfaces that have been treated in this way show increased adhesion^{18–20}, it seems that the effect may originate in relatively poor adhesion between the metal and the polymer substrate. Results similar to those observed for plasma-treated HDPE were obtained with amorphous, atactic polystyrene (a-PS) films coated with 50- and 100-nm gold, when embossing was conducted at 80–100 °C, that is, around its T_g .

We used 50-nm gold/FEP and a-PS bilayers embossed above T_m and T_g of the polymer substrates (at 330 and 200 °C, respectively) as references. No microcutting of the metal layer occurred, only bending and deep-drawing. Clearly, the polymer substrate must be able to sustain without substantial flow the load necessary for indentation of the metal, and fixate the metal layer to enable cutting and prevent deep-drawing. Only after the metal layer has been pierced should plastic flow of the substrate begin, which thus needs to be in a plastically deformable solid state and not in either the liquid or rigid (brittle) glass phase during microcutting.

The following factors seem to control events during microcutting of metal/polymer bilayers. Applying a load on the master positioned onto the bilayer will initially cause local plastic deformation below its wedge tips and bending of the metal layer. Its deflection, which

depends on the load necessary for indentation (for a wedge of angle 70°, about 1.4 times the yield stress²¹), and the elastic moduli of both the substrate and the metal cause local shear stresses at the metal/polymer interface. If these shear stresses exceed the shear strength of the interface or that of the substrate, delamination will occur²² and the substrate no longer fixates the metal layer. Consequently, cutting ceases in favour of deep-drawing of the metal layer, or is accompanied by it (compare the ‘doubling’ effect). We observed that ‘doubling’ is favoured for thicker films (increased interfacial shear stress) and lower adhesion (decreased interfacial shear strength). Regarding indentation of the metal layer, here we describe a simple method of identifying some limits associated with the microcutting process. When a rigid, frictionless wedge penetrates a body, matter is squeezed towards the surface forming a lip next to the wedge. As the material can flow towards the free surface during indentation, the effects of elastic compressibility will be small compared to those of plastic deformation, and can be neglected. A plane-strain analysis of the relevant mechanics²¹, assuming perfectly rigid-plastic behaviour, showed that the shape of the plastic-deformation zone depends only on the wedge top angle β , which, in turn, determines the centred fan angle, α :

$$\cos(2\beta - \alpha) = \cos(\alpha)/[1 + \sin(\alpha)] \quad (1)$$

The length of the displaced surface, h , is:

$$h/k = [\sin(\beta) + \cos(\beta - \alpha)]/\cos(\alpha) \quad (2)$$

Figure 3 is the unity diagram of this zone in which all distances are scaled to the wedge-indentation depth k . When neighbouring plastic-deformation zones overlap, further indentation may cause plastic deformation of the entire metal layer, which could limit resolution of our technique. Using equations (1) and (2), we calculate that for the present wedge angle of 70°, $\alpha = 22^\circ$ and $h/k \approx 1$, from which follows that for a 50-nm gold layer the smallest possible lateral ‘wire’ dimension is about 100 nm; for this, an inter-wedge spacing of about 170 nm would be required. As α depends only on β , the shape of the plastic zone will not change after the metal layer has been pierced and indentation of the polymer substrate commences. Thus, for frictionless indentation, as long as both bilayer components are ductile, their detailed mechanical properties are irrelevant for the final (microcut) geometry; this holds promise for the broad applicability of microcutting, also for other material combinations. In the probable case that friction accompanies indentation, additional forces may act and force the metal lips downwards²³.

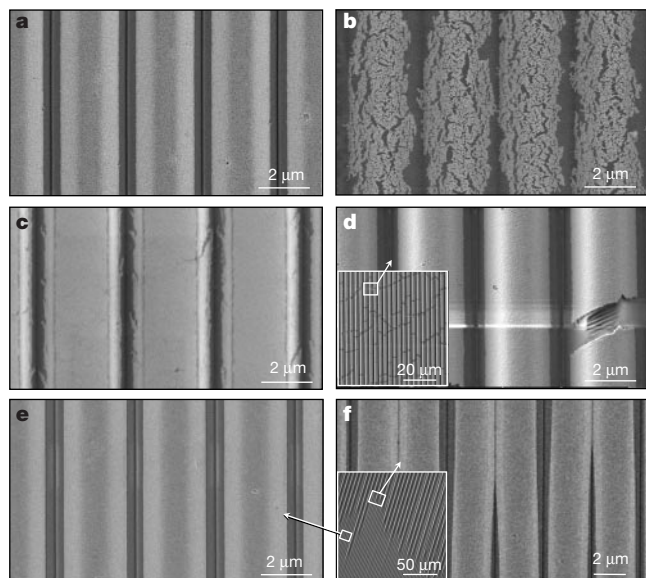


Figure 2 ESE micrographs of microcut metal/poly(tetrafluoroethylene-co-hexafluoropropylene) (FEP) bilayer structures revealing various aspects of the process (a–f). **a**, A microcut 50-nm gold/FEP bilayer. Photomicrograph **b** shows that cutting of a 20-nm gold/FEP substrate resulted in breaking-up of the metal layer, which was substantially less pronounced for 80:20 per cent by weight gold–palladium of the same thickness (**c**). Microcut structures of the latter alloy, unlike gold, displayed cracks when the layer thickness was greater than 50 nm (**d**). **e**, Micrograph of an embossed bilayer film of 100 nm gold/FEP where the metal layer is homogeneously patterned, and a transition area which displays microcutting in every second groove (**f**). Inset, low-magnification optical micrograph of that same region.

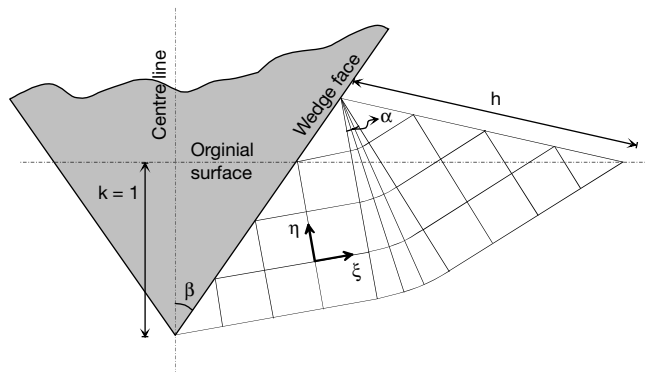


Figure 3 Diagram of a wedge (grey) penetrating without friction a rigid, perfectly plastic body after ref. 21. Unity diagram (that is, wedge indentation depth $k = 1$) of the plastic region for a total wedge angle of 70° ($\beta = 35^\circ$), leading to a centred fan angle $\alpha = 22^\circ$. The orthogonal curvilinear coordinates ξ and η , also called slip-lines, denote the directions of maximum shear stress, and are, therefore, required to make a 45° angle with both the wedge face and the displaced (free) surface.

Furthermore, if the master is fully pressed into the bilayer, the indented metal may bend to adopt the inter-wedge geometry. If the metal-grain size is sufficiently small, the bending could allow further reduction of the lower dimension limit of the metal features, evidently at the expense of increased separation.

To illustrate its potential to produce microelectrode arrays, aluminium-metallized poly(ethylene terephthalate) (PET) films were microcut at 90°C, yielding regularly spaced aluminium 'micro-wires' on the polymer substrate. Electrical measurements indicated a resistance parallel to the micro-wires that was of the same order of magnitude (about 30 Ω) as that of the unembossed aluminium coating (about 10 Ω) and essentially identical to that of the insulating PET film perpendicular to them.

Some optical characteristics of microcut metal/polymer bilayers are presented in Fig. 4. Figure 4a shows infrared transmittances T measured without an auxiliary polarizer: of a patterned 20-nm

aluminium/HDPE bilayer with embossed grooves positioned vertically, T_v , and horizontally, T_h , with respect to the instrument plane; the transmittance of two crossed films, T_{hv} ; and that of a reference, non-metallized, embossed HDPE film, T_0 . The following values are also shown as functions of wavelength λ , for a wire grid of metal strips of width a and spacing d (refs 3, 24): the calculated²⁵ experimental polarizing efficiency, $PE^{exp} = \sqrt{1 - [4T_0T_{hv}/(T_h + T_v)^2]}$; single-piece transmittance, $T_{sp}^{exp} = T_h + T_v/2T_0$; and their corresponding theoretical values PE^{th} and T_{sp}^{th} (corrected for T_0).

$$PE^{th} = (T_{\perp}^{th} - T_{\parallel}^{th}) / (T_{\perp}^{th} + T_{\parallel}^{th})$$

$$T_{sp}^{th} = \frac{1}{2} (T_{\perp}^{th} + T_{\parallel}^{th})$$

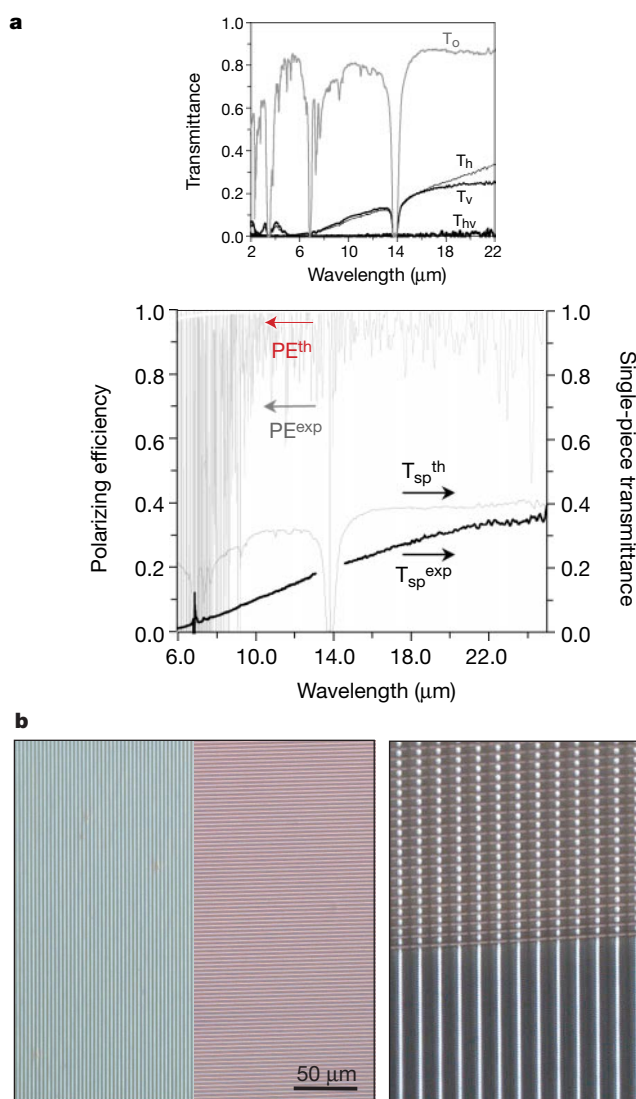


Figure 4 Illustrations of various optical properties of microcut metal/polymer films. **a**, Top, Infrared transmittance as function of wavelength of patterned 20-nm aluminium/high-density polyethylene (HDPE) bilayer structures with grooves vertical, T_v , and horizontal, T_h , to the instrument plane, and transmittance of two crossed films, T_{hv} , and of a reference, embossed HDPE film, T_0 . Bottom, experimental polarizing efficiency PE^{exp} and single-piece transmittance T_{sp}^{exp} and their calculated values PE^{th} and T_{sp}^{th} (see text). **b**, Optical micrographs of patterned bilayer structures of 100-nm gold/poly(vinyl alcohol) (PVA). Left,

single-embossed films appeared red with the imprinted grooves oriented perpendicular, and green when parallel to the polarization direction of the incident light. Right, partially double-embossed film of which one area was cut in every second groove in the first patterning step (vertical lines); in the second step the master was rotated by about 90° and the gold layer was cut in every groove, resulting in a regular dot pattern of the metal.

$$T_{\perp}^{\text{th}} = \frac{4nA^2}{1 + A^2(1+n)^2}; \quad T_{\parallel}^{\text{th}} = \frac{4nB^2}{1 + (1+n)^2}B^2$$

$$\frac{1}{A} = \frac{4d}{\lambda} \left(\ln \left[\operatorname{cosec} \frac{\pi(d-a)}{2d} \right] + \frac{Q_2 \cos^4 \left[\frac{\pi(d-a)}{2d} \right]}{1 + Q_2 \sin^4 \left[\frac{\pi(d-a)}{2d} \right]} \right) + \frac{1}{16} \left(\frac{d}{\lambda} \right)^2 \left(1 - 3 \sin^2 \left[\frac{\pi(d-a)}{2d} \right] \right)^2 \cos^4 \left[\frac{\pi(d-a)}{2d} \right]$$

$$B = \frac{d}{\lambda} \left(\ln \left[\operatorname{cosec} \frac{\pi a}{2d} \right] + \frac{Q_2 \cos^4 \left[\frac{\pi a}{2d} \right]}{1 + Q_2 \sin^4 \left[\frac{\pi a}{2d} \right]} \right) + \frac{1}{16} \left(\frac{d}{\lambda} \right)^2 \left(1 - 3 \sin^2 \left[\frac{\pi a}{2d} \right] \right)^2 \cos^4 \left[\frac{\pi a}{2d} \right]$$

$$Q_2 = \frac{1}{\left[1 - \left(\frac{d}{\lambda} \right)^2 \right]^2} - 1 \quad (3)$$

Here $T_{\perp}^{\text{th}}$ and $T_{\parallel}^{\text{th}}$ are the theoretical transmittances for radiation polarized perpendicular and parallel to the grid wires, respectively, and n is the refractive index of the substrate (here 1.51). Figure 4a shows that at large wavelengths ($>20 \mu\text{m}$), where the microcut metal/polymer bilayer acts as a zero-order grating, both the experimental and theoretical polarizing efficiencies are approximately 1.0 and single-piece transmittances approach 0.4. This illustrates the potential use of the present systems as highly efficient infrared polarizers (commonly, values of $\text{PE} > 0.95$ and $T_{\text{sp}} \approx 0.35$ are required²⁶).

Figure 4b (left) shows another interesting optical property of microcut metal/polymer films: patterned 100-nm gold/poly(vinyl-alcohol) (PVA) bilayers displayed pronounced polarization-dependent colours; that is, they appeared red or green when examined with the embossed grooves perpendicular or parallel to the polarizing direction of incident light. This has been previously noticed for oriented metal-nanoparticle/polymer composites and shown to allow generation of multiple colours in single electro-optical elements²⁷.

Multiple microcutting can also be carried out. Figure 4b (right) is an optical micrograph of a doubly embossed 100-nm gold/PVA film, which shows that with simple masters, relatively complex structures may be produced. We can readily envision other useful extensions of the present process, such as microcutting of multiple-layer films. Adding a simple dissolution step of the polymer substrate would enable fabrication of free-standing metal structures when appropriate cutting patterns are used. □

Methods

Materials used in this work were purchased from the following sources: poly(tetrafluoroethylene-co-hexafluoropropylene), $T_m = 260^\circ\text{C}$, $T_g = 80^\circ\text{C}$ (Teflon FEP 100), from DuPont; high-density polyethylene (HDPE), $T_m = 135^\circ\text{C}$, $T_g \approx -120^\circ\text{C}$, average molecular weight, $M_w \approx 90,000 \text{ g mol}^{-1}$ (Stamylan HD 7048), from DSM; atactic polystyrene, (a-PS), $T_g \approx 95^\circ\text{C}$, $M_w \approx 230,000 \text{ g mol}^{-1}$, from Aldrich; poly(ethylene terephthalate) (PET) films (thickness $50 \mu\text{m}$), $T_m = 260^\circ\text{C}$, $T_g \approx 80^\circ\text{C}$, metallized with 35-nm aluminium, from Goodfellow; poly(vinyl alcohol) (PVA), $T_m = 235^\circ\text{C}$, $T_g \approx 85^\circ\text{C}$, $M_w \approx 100,000 \text{ g mol}^{-1}$, from Aldrich; gold 99.99% and aluminium 99.9%, from Balzers; 80:20 per cent by weight gold–palladium alloy, from BAL-TEC. Polymer films of thicknesses in the range of 125–175 μm were prepared by standard melt-compression

moulding, except those of PVA, which were cast from a 1 per cent by weight solution in deionized water and subsequently dried for 48 h at 50°C , leaving films of about 90 μm thickness. Metal films were evaporated employing a MED-020 instrument (BAL-TEC) operated at 1×10^{-5} mbar (gold and gold–palladium) or a BAE 370 coating system (Balzers) operated at 1.3×10^{-6} mbar (aluminium). The master for embossing was an ultrasharp calibration grating TGG01 for scanning probe microscopy (SPM) scanners, area 9 mm^2 , from NT-MTD, Moscow. Oxygen-plasma treatments were carried out with a plasma cleaner (PDC-32G) from Harrick Scientific, operated at 40 W with a water-enriched oxygen feed. Embossing was performed under the following conditions: gold/FEP and 80:20 per cent by weight gold–palladium/FEP, pressure on the master of 850 g mm^{-2} for 5 min at 200°C ; gold/HDPE, same conditions but at 100°C ; gold/PVA, same conditions but at 150°C ; gold/a-PS and aluminium/PET, 1 kg mm^{-2} for 1 h at $80-100^\circ\text{C}$. Electrical conductivities were measured with a Metex M-3610D digital multimeter, CE-BIT ELEKTRONIK. Cross-sections of metal-coated polymer films were prepared by embedding the multilayer films in a pre-polymeric liquid, consisting of 1 ml monomer mixture (41.3 wt% Epon 812, 54 wt% dibutylphthalate, 4.7 wt% Durcupan ACM), 1 ml hardener (DDSA) and five drops of accelerator (DMP30), all from Fluka, which was subsequently cured at 60°C for 24 h; sectioning was performed with a Leica RM 2155 microtome.

Received 12 February; accepted 14 August 2000.

- Rogers, J. A., Bao, Z. & Raju, V. R. Nonphotolithographic fabrication of organic transistors with micron feature sizes. *Appl. Phys. Lett.* **72**, 2716–2718 (1998).
- Gale, M. T., Kane, J. & Knop, K. ZOD images: Embossable surface-relief structures for color and black-and-white reproduction. *J. Appl. Photo. Eng.* **4**, 41–47 (1978).
- Auton, J. P. Infrared transmission polarizers by photolithography. *Appl. Opt.* **6**, 1023–1027 (1967).
- Chan, V. Z.-H. *et al.* Ordered bicontinuous nanoporous and nanorelief ceramic films from self-assembling polymer precursors. *Science* **286**, 1716–1719 (1999).
- Gau, H., Herminghaus, S., Lenz, P. & Lipowsky, R. Liquid morphologies on structured surfaces: From microchannels to microchips. *Science* **283**, 46–49 (1999).
- Kataoka, D. E. & Troian, S. M. Patterning liquid flow on the microscopic scale. *Nature* **402**, 794–797 (1999).
- Moreau, W. M. *Semiconductor Lithography* (Plenum, New York, 1988).
- Madou, M. *Fundamentals of Microfabrication* (CRC Press, Boca Raton, 1997).
- Kumar, A. & Whitesides, G. M. Features of gold having micrometer to centimeter dimensions can be formed through a combination of stamping with an elastomeric stamp and an alkanethiol “ink” followed by chemical etching. *Appl. Phys. Lett.* **63**, 2002–2004 (1993).
- Gale, M. T. Replication technology for holograms and diffractive optical elements. *J. Imag. Sci.* **41**, 211–220 (1997).
- Stutzmann, N., Tervoort, T. A., Bastiaansen, C., Feldmann, K. & Smith, P. Solid-state replication of relief structures in semi-crystalline polymers. *Adv. Mater.* **12**, 557–562 (2000).
- Kim, E., Xia, Y. & Whitesides, G. M. Polymer microstructures formed by moulding in capillaries. *Nature* **376**, 581–584 (1995).
- Kim, C., Burrows, P. E. & Forrest, S. R. Micropatterning of organic electronic devices by cold-welding. *Science* **288**, 831–833 (2000).
- Echlin, P. Recent advances in specimen coating techniques. *Scanning Electron Microsc.* **14**, 79–90 (1981).
- Bråten, T. High resolution scanning electron microscopy in biology: Artefacts caused by the nature and mode of application of the coating material. *J. Microsc.* **113**, 53–59 (1978).
- Petch, N. J. The cleavage strength of polycrystals. *J. Iron Steel Inst.* **174**, 25–28 (1953).
- Hall, E. O. The deformation and ageing of mild steel: III Discussion of results. *Proc. Phys. Soc. Lond. B* **64**, 747–753 (1951).
- Gerenser, L. J. in *Metallization of Polymers* (eds Sacher, E., Pireaux, J.-J. & Kowalczyk, S. P.) 433–452 (American Chemical Society, Washington, 1990).
- Gerenser, L. J. in *Metallized Plastics: Fundamentals and Applications* (ed. Mittal, K. L.) 113–128 (Marcel Dekker, New York, 1993).
- Sacher, E. in *Metallization of Polymers* (eds Sacher, E., Pireaux, J.-J. & Kowalczyk, S. P.) 1–7 (American Chemical Society, Washington, 1990).
- Hill, R., Lee, E. H. & Tupper, S. J. The theory of wedge indentation of ductile materials. *Proc. R. Soc. A* **188**, 273–289 (1947).
- Chechenin, N. G., Bottiger, J. & Krog, J. P. Nanoindentation of amorphous aluminium-oxide films II. Critical parameters for the breakthrough and a membrane effect in thin hard films on soft substrates. *Thin Solid Films* **261**, 228–235 (1995).
- Pivin, J. C., Lebouvier, D., Pollock, H. M. & Felder, E. Fields of plastic deformation in indented bilayers: Comparison between kinematic calculations and experimental data obtained at scales ranging from one centimetre to ten nanometres. *J. Phys. D* **22**, 1443–1450 (1989).
- Bennett, J. M. & Bennett, H. E. in *Handbook of Optics* (eds Driscoll, W. G. & Vaughan, W.) Ch. 10 (McGraw-Hill, New York, 1978).
- Dirix, Y., Tervoort, T. A. & Bastiaansen, C. Optical properties of oriented polymer/dye polarizers. *Macromolecules* **28**, 586–491 (1995).
- Dirix, Y., Tervoort, T. A. & Bastiaansen, C. Optical properties of oriented polymer/dye polarizers. 2. Ultimate properties. *Macromolecules* **30**, 2175–2177 (1997).
- Dirix, Y., Bastiaansen, C., Caseri, W. & Smith, P. Oriented pearl-necklace arrays of metallic nanoparticles in polymers: A new route towards polarization-dependent color filters. *Adv. Mater.* **11**, 223–227 (1999).

Acknowledgements

We thank P. Schmit for performing environmental scanning electron microscopy and L. Fall for assistance with infrared spectroscopy measurements.

Correspondence and requests for materials should be addressed to T.A.T. (e-mail: tervoort@ifp.mat.ethz.ch).

Quantifying the uncertainty in forecasts of anthropogenic climate change

Myles R. Allen*, Peter A. Stott†, John F. B. Mitchell‡, Reiner Schnur‡ & Thomas L. Delworth§

* Space Science and Technology Department, Rutherford Appleton Laboratory, Chilton, Didcot OX11 0QX, UK

† Hadley Centre for Climate Prediction and Research, The Meteorological Office, London Road, Bracknell RG12 2SZ, UK

‡ Max-Planck-Institut für Meteorologie, Bundesstrasse 55, 20146, Hamburg, Germany

§ Geophysical Fluid Dynamics Laboratory/NOAA, PO Box 308, Princeton, New Jersey 08542-0308, USA

Forecasts of climate change are inevitably uncertain. It is therefore essential to quantify the risk of significant departures from the predicted response to a given emission scenario. Previous analyses of this risk have been based either on expert opinion¹, perturbation analysis of simplified climate models^{2–5} or the comparison of predictions from general circulation models⁶. Recent observed changes that appear to be attributable to human influence^{7–12} provide a powerful constraint on the uncertainties in multi-decadal forecasts. Here we assess the range of warming rates over the coming 50 years that are consistent with the observed near-surface temperature record as well as with the overall patterns of response predicted by several general circulation models. We expect global mean temperatures in the decade 2036–46 to be 1–2.5 K warmer than in pre-industrial times under a ‘business as usual’ emission scenario. This range is relatively robust to errors in the models’ climate sensitivity, rate of oceanic heat uptake or global response to sulphate aerosols as long as these errors are persistent over time. Substantial changes in the current balance of greenhouse warming and sulphate aerosol cooling would, however, increase the uncertainty. Unlike 50-year warming rates, the final equilibrium warming after the atmospheric composition stabilizes remains very uncertain, despite the evidence provided by the emerging signal.

Many attempts have been made to forecast the response of the climate system to anthropogenic changes in atmospheric composition. For example, the first five squares in Fig. 1 (from left to right) show predicted global mean temperature in the decade 2036–46 relative to the pre-industrial (control) climate for four atmosphere–ocean general circulation models (A–OGCMs)^{12–15} all driven with approximately the same scenario (IS92a, ref. 16) of greenhouse gas and sulphate aerosol (GS) concentrations. (The fourth square is from a model that includes the effects of indirect sulphate and tropospheric ozone concentrations¹⁴.)

A basic problem with all such predictions to date has been the difficulty of providing any systematic estimate of uncertainty. Predictions in Fig. 1 range from 1.1 to 2.3 K, but translating inter-model spread into an objective uncertainty range is problematic because these models do not necessarily span the full range of known climate system behaviour. Their climate sensitivities (equilibrium warming on doubling carbon dioxide), for example, all lie in the range 2.5–3.5 K: a smaller range than even the most optimistic current estimate of uncertainty in this parameter. Other A–OGCMs exist with sensitivities outside this range, but a probabilistic interpretation of the spread of ‘current’ predictions is inevitably subjective.

An alternative approach is illustrated by the heavy solid line in Fig. 2. This shows how the predicted warming by 2036–46 under the IS92a GS scenario varies with the simulated rate of anthropogenic

warming over the twentieth century as we vary the prescribed climate sensitivity in a simple climate model². The plot is close to a straight line through zero, indicating a simple linear relationship between the amplitude of the signal observed to date and the size of mid-twenty-first-century warming. Almost the same relationship emerges if we assume a different rate of oceanic heat uptake¹⁷ (heavy dotted line). Thus, if the range of twentieth-century warming trends attributable to anthropogenic influence were 0.25–0.5 K per century, this model suggests that the uncertainty range in 2040s temperatures would be 1–2 K, irrespective of the accuracy of the model’s climate sensitivity or rate of oceanic heat uptake.

In contrast, the relationship between the observed signal and the equilibrium climate sensitivity is both nonlinear and dependent on the assumed rate of oceanic heat uptake. A range in recent anthropogenic warming rates of 0.25–0.5 K per century would translate into an uncertainty in sensitivity of 1.5–4 K if we assume the faster rate of oceanic heat uptake (thin dotted line in Fig. 2, right-hand scale). With the slower rate (thin solid line), no useful upper bound could be placed on the climate sensitivity on the basis of twentieth-century temperature trends. Thus we cannot estimate climate sensitivity from recent observed surface temperature trends without an independent estimate of the rate of oceanic response^{2,5,17}. Even if this response time were known, if it turns out to be towards the slower end of the current uncertainty range, then it may still be impossible to provide a useful upper bound on climate sensitivity on the basis of recent trends.

We expect similar relationships to those in Fig. 2 to hold in more complex systems, provided both the strength of atmospheric feedbacks and the rate of oceanic heat uptake do not change in response to perturbations of this magnitude. The fact that most A–OGCMs give an almost linear response to a linear increase in radiative forcing⁶ provides support for this assumption, but direct perturbation analysis of more complex models with realistic forcing trajectories is required to explore fully its validity¹⁸.

Considering only the transient response, the problem now becomes: what fraction of the recent observed warming should be attributed to anthropogenic influence? Climate change detection

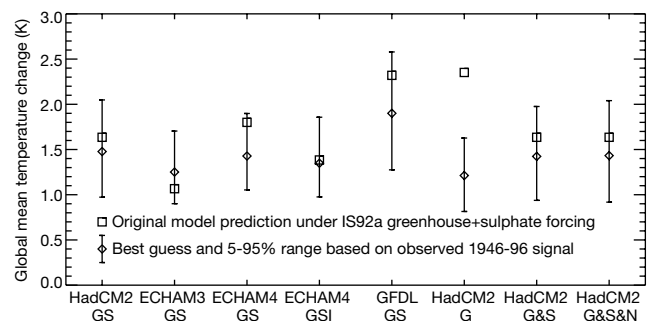


Figure 1 Predicted anthropogenic warming by about 2040 under the IS92a scenario before and after reconciling model simulations with the recent climate record. Squares, temperature change relative to pre-industrial (control) climate in a range of climate models for the decade 2036–46 under the IS92a scenario of greenhouse gas and sulphate (GS) forcing (greenhouse gas only in sixth case, G). Diamonds, scaled temperature change after reconciling model-simulated large-scale patterns of near-surface temperature anomalies (taken about the 1896–1996 mean) over the five decades 1946–96 with the corresponding observed signal using an optimal fingerprint algorithm. First six cases, scaling single GS or G simulation; seventh case, net anthropogenic warming estimating greenhouse gas and sulphate responses independently; eighth case, net anthropogenic warming estimating greenhouse gas, sulphate and natural (combined solar and volcanic) signals independently, assuming no naturally forced change after 1996. Vertical bars, uncertainty in the scaled response based on uncertainty in the scaling factor(s) required to reconcile the models with observed changes.

techniques^{19,20} provide an estimate of this fraction or, more specifically, an estimate of the factor by which we have to scale a model-simulated trajectory to match the magnitude of the observed anthropogenic signal²⁰, together with an objective estimate of the corresponding range of uncertainty.

The first five diamonds in Fig. 1 show the various models' predictions for total warming by 2036–46 under the IS92a GS scenario after scaling their respective trajectories to match the amplitude of observed near-surface temperature anomalies (relative to the 1896–1996 mean) over the 1946–96 period. By the mid-twenty-first century, sampling uncertainty in the raw model predictions is small relative to the climate change signal, particularly if ensemble simulations are available. Thus a factor-of-two uncertainty in the scaling required on the model-simulated twentieth-century response translates into a factor-of-two uncertainty in the scaled prediction.

Rather than simply using temperature trends as a measure of the strength of the observed signal, we use a full spatio-temporal "fingerprint"¹⁹ of the various responses^{11,21}. This provides an "optimal estimate"¹⁰ of the scaling required to match the amplitude of the observed signal²⁰, minimizing uncertainty due to internal climate variability. Application of this scaling improves inter-model consistency, with predictions from un-responsive models being scaled up and predictions from highly responsive models scaled down. The GFDL model prediction remains higher than the others, implying that the A–OGCMs are diverging by more than we would expect if they were simply providing scaled versions of the same underlying trajectory. The further we predict into the future, the greater this divergence is likely to be, so our approach is only valid over timescales up to the length of the observational record used. Estimated uncertainties in the amplitude of the scaled response, based on the individual models' estimates of internal climate variability, are shown by the vertical bars.

To show what happens if an important process is omitted, the sixth square in Fig. 1 shows the predicted warming by 2036–46 from the HadCM2 model forced with rising greenhouse gases alone. The

raw prediction is over half a degree warmer than the same model under GS forcing, but the fingerprint analysis suggests that this model trajectory would need to be scaled down significantly to be consistent with recent observations. After such scaling, the resulting best-guess and uncertainty range is in better agreement with the GS simulations. This point is important because other processes may also have been omitted from all the GS simulations. Provided these processes have a proportionally similar impact on the signal observed to date as on early twenty-first-century warming (as would be the case for an atmospheric feedback that scales approximately linearly with the surface temperature change), their omission from the models would not affect the estimated scaled prediction.

Errors which only show themselves in the future, such as a failure to represent a sudden shut-down in the thermohaline circulation, would not be accounted for in this analysis. Most currently simulated circulation changes seem to be relatively gradual over the timescales of interest²², but the possibility of sudden nonlinear climate change limits the forecast lead time over which this approach can be pursued. The assumption that the spatio-temporal patterns of response are independent of the response amplitude appears to be acceptable for large-scale surface temperature changes⁸, but it would not be valid for changes in precipitation²³ or atmospheric circulation²⁴, or for cases in which the forcing changes abruptly over the period of interest.

Another assumption underlying the first five bars in Fig. 1 is that the relative amplitude of the responses to greenhouse gases and to sulphate aerosols is as simulated in these GS experiments, so the combined response can be represented in each case by a single spatio-temporal pattern. Given the considerable uncertainty in the sulphate forcing and response, this is difficult to justify. For the HadCM2 climate model, we have separate multi-member ensemble simulations of the greenhouse-gas-only (G) and the GS response. The fingerprinting approach allows separate estimates of the amplitude of both the greenhouse and sulphate signals in the observations, together with their joint uncertainty range^{9,19,20}. By scaling

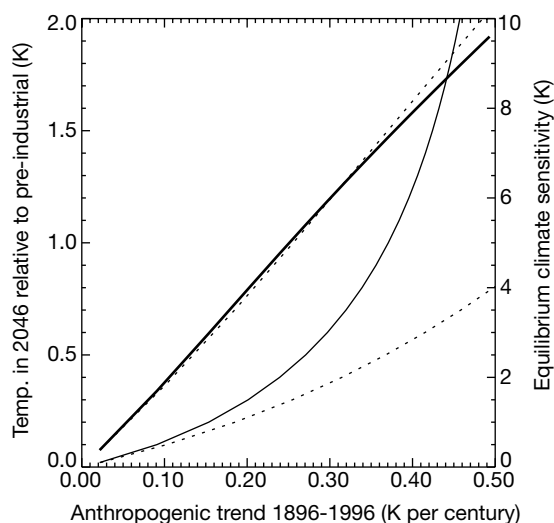


Figure 2 Transfer functions relating past and future climate change in a simple climate model. Heavy solid line, left-hand scale: relationship between simulated global mean temperature trends over the period 1896–1996 and the predicted total anthropogenic warming by the decade 2036–46, obtained by varying the climate sensitivity in a simple climate model² under IS92a greenhouse and sulphate forcing. Heavy dotted line, relationship obtained assuming a different rate of oceanic heat uptake (effective vertical diffusivity of $0.25 \text{ m}^2 \text{ s}^{-1}$ versus $2.0 \text{ m}^2 \text{ s}^{-1}$ in the base case). Light solid and dotted curves, right hand scale: corresponding relationships between simulated 1896–1996 temperature trends and equilibrium climate sensitivity.

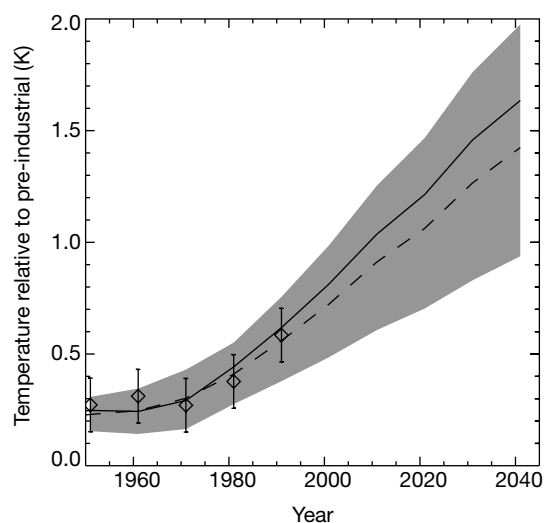


Figure 3 Global temperatures under IS92a consistent with the recent climate record. This figure shows simulation and forecast of decadal global mean temperatures relative to pre-industrial (control) values as predicted by HadCM2 under the IS92a GS scenario. Solid line, mean of original 4-member ensemble simulation. Dashed line, after scaling the model-simulated spatio-temporal patterns of response to greenhouse gas and sulphate forcing individually to give the best combined fit to the observations over the 1946–96 period. Shaded band, 5–95% confidence interval on scaled response. Diamonds, observed decadal global mean temperatures (anomalies about the 1896–1996 mean, added for display to the corresponding mean of the GS ensemble); vertical bars, ± 2 s.d. of decadal mean temperatures from HadCM2 control.

their individual contributions to model-predicted future warming accordingly, we arrive at an uncertainty estimate which does not depend on the amplitude of the model-simulated response to either forcing agent.

The seventh bar in Fig. 1 shows the resulting range of uncertainty in the HadCM2 prediction for 2036–46, while Fig. 3 shows how this uncertainty range evolves over time. The solid line shows the original model prediction; the dashed line and shaded band show the median and 5–95% uncertainty range after scaling the simulated greenhouse and sulphate signals to match observed large-scale temperatures over the 1946–96 period. The observed spatially-averaged global mean temperatures of the five decades 1946–1996 are shown (diamonds) for illustration only: scaling factors are estimated from the full spatio-temporal pattern of temperature change. The figure shows the range of uncertainty in the underlying anthropogenic trend: superimposed on this would be uncertainty due to internal variability, indicated by the vertical bars on the observations, and due to other forcing agents.

Estimating separately the amplitudes of greenhouse and sulphate signals does not appear to increase forecast uncertainty: the first and seventh bars in Fig. 1 are almost identical. The reason is that the combination of greenhouse warming and sulphate cooling predicted under the IS92a emissions scenario happens to be particularly well constrained by the observed signal at the global level. This is shown in Fig. 4, which focuses on warming over the coming 50 years rather than warming relative to pre-industrial times as shown in Figs 1–3. The dotted contours show how projected warming over the five decades to 2046 depends on the scaling factors applied to the model-simulated greenhouse and sulphate signals. The raw model prediction is a greenhouse warming of 1.35 K and a sulphate cooling of 0.35 K over this period, giving a net warming of 1 K assuming both scaling factors are unity (the square). If we scale the sulphate signal up or down by a factor of 1.35/0.35 faster than we scale the

greenhouse signal (that is, moving along an isoline of future warming), the net warming 1996–2046 is unchanged.

The cross and shaded region show the best guess and joint 90% uncertainty range on the scaling factors required on the model-simulated greenhouse and sulphate response-patterns to reproduce observed temperatures over the 1946–96 period. The uncertainty range is strongly tilted, meaning the model could over- (or under-) estimate the magnitude of recent greenhouse warming and still be consistent with the observed signal provided it also over- (under-) estimates the magnitude of sulphate cooling. The principal axis of the uncertainty region is close to parallel to the isolines of future warming, so uncertainty in predicted 50-year warming under the IS92a scenario is relatively low (the best-guess of 0.9 K and the 5–95% range of 0.55–1.2 K are shown by the diamond and the solid bar).

Any reduction in future sulphate cooling not only increases the best-guess net future warming, but it also substantially increases the uncertainty range consistent with current observations. The reason is that the two-dimensional confidence region would no longer be so well aligned with the isolines of predicted warming. For example, if the predicted 0.35 K cooling due to sulphates over the 1996–2046 period is eliminated altogether (making the contours in Fig. 4 vertical), the best-guess net warming over this period increases by 0.2 K, or about 20% (less than 0.35 K because the best-guess scaling on the sulphate signal in Fig. 4 is 0.6), whereas the upper bound on the 5–95% range increases by almost 50% to 1.7 K. This would also be the case for other factors, such as stratospheric ozone depletion, whose influence on climate is less easy to detect than the agents considered here²⁰, but in which trends are also expected to reverse over the next few years.

Thus far, we have assumed that the observed record consists only of anthropogenic signals and internal variability. Although this assumption is consistent with the available data^{11,20,25}, we have reason to expect that natural external factors have also affected temperatures over the twentieth century^{10,11,26}. If we include the combined response (as estimated by HadCM2¹¹) to solar variability and volcanic aerosols in a three-way fingerprint analysis²¹, the uncertainty range in projected anthropogenic warming (eighth bar in Fig. 1) is almost unchanged. This is because the inclusion of this estimate of the natural signal does not have a detectable impact on the estimated amplitude of the anthropogenic signal (the use of decadal mean data minimizes the impact of individual volcanic eruptions and the 11-year solar cycle). Nevertheless, uncertainty in the fraction of recent warming attributable to natural versus anthropogenic influences, together with uncertainty in future natural forcing, remain important caveats.

We have focused on uncertainty in the climate system's response to a given profile of future greenhouse gas concentrations: uncertainty in other sources and sinks²⁷ and in future emissions²⁸ are also important. For example, the spread of responses by 2046 of a single climate model to the full range of scenarios considered by the IPCC Special Report on Emissions Scenarios (SRES)²⁸ is almost as large as the uncertainty in the response to a single scenario shown in Fig. 3. Not all of the SRES emissions scenarios, however, are based on explicit social and economic models²⁹, so it remains debatable whether uncertainty in future emissions is yet comparable to uncertainty in the response. Nevertheless, a quantitative assessment of the range of possible responses to a given emission path will increasingly be required to replace the best-guess projections hitherto provided by A-OGCMs. The approach described here provides a means of incorporating information on observed climate change into these ranges, until full-scale ensemble climate prediction systems become available¹⁸. A 5–95% range of 1–2.5 K above pre-industrial values by the 2040s under the IS92a scenario represents a first assessment, subject to various assumptions. We can expect this range to increase as these assumptions are relaxed in future work, but we can also expect it to be reduced if the observed

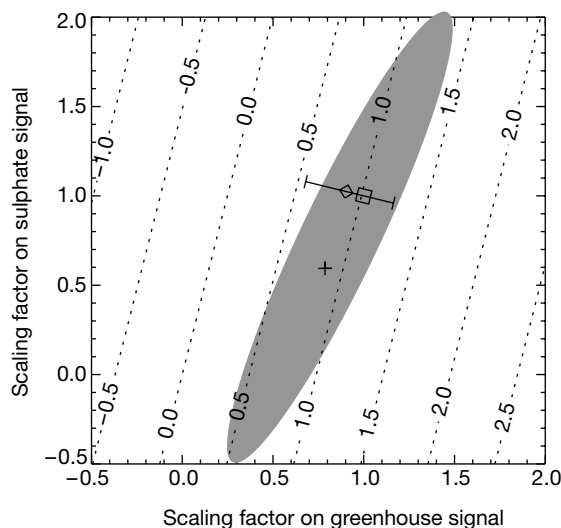


Figure 4 Forecast anthropogenic warming 1996 to 2046 under IS92a. Dotted contours, isolines of global mean warming (in K) between the decade 1986–96 and the decade 2036–46 as a function of the scaling factors applied to the raw HadCM2 prediction of a 1 K warming (square). Cross and shaded region, best-guess and joint 90% confidence region on estimated scaling factors required to reproduce observed large-scale near-surface temperatures over the 1946–96 period using response-patterns simulated by HadCM2. Diamond and solid error bar, best-guess and 5–95% range on forecast warming (read off from the contours) obtained by a probability-weighted sum of ‘allowed’ scaling factors along the isolines of future warming. The error bar is superimposed on the raw prediction (square) for comparison with Fig. 1: provided the diamond and cross lie on the same isoline of future warming, its location is otherwise arbitrary.

climate change signal continues to strengthen over the next few years. □

Methods

HadCM2-GS¹⁶ and GFDL-GS¹⁵ simulations are based on 4- and 5-member ensembles respectively forced with observed greenhouse gas and parametrized direct sulphate forcing to 1990 followed by 1% yr⁻¹ compound increase in CO₂ (close to the IS92a scenario in terms of radiative forcing²³) and IS92a projected sulphate loadings. ECHAM3-GS two-member ensemble¹³ and ECHAM4-GS single simulation¹⁴ are both based on observations followed by IS92a; the ECHAM4-GS1 single stimulation includes the impact of indirect sulphate forcing and tropospheric ozone changes. Model–observation comparison is based on decadal mean near-surface temperatures over the period 1946–96. Data are expressed as departures from the 1896–1996 mean, exploiting the fact that recent decades have been generally warmer than the preceding half-century without attempting to fit the details of surface temperature changes in poorly-sampled earlier decades. Ensemble members and 100-year segments of the control were masked with the pattern of missing data in the observations before computing means and anomalies, filtered to retain only scales greater than 5,000 km (refs 11, 30) and projected onto the 10 leading modes of internal spatio-temporal variability of the individual model control simulations (ECHAM3 control used for ECHAM4). Scaling factors (diamond/square ratios in Fig. 3) are estimated using standard optimal fingerprinting¹⁹ modified to account for the presence of sampling noise in model-simulated signals^{31,32}. Uncertainty estimates reflect uncertainty in scaling factors given interdecadal variability in the individual model control simulations (HadCM2: 1,700 years, ECHAM3: 1,900 years, GFDL: 1,000 years). In each case the first half of the control was used to define the detection space and for optimization, the second for uncertainty analysis.

Received 30 December 1999; accepted 2 August 2000.

- Morgan, M. G. & Keith, D. W. Subjective judgements by climate experts. *Environ. Policy Anal.* **29**, 468–476 (1995).
- Hansen, J. *et al.* Climate response times: dependence on climate sensitivity and ocean mixing. *Science* **229**, 857–859 (1985).
- Raper, S. C. B., Wigley, T. M. L. & Warrick, R. A. in *Rising Sea Level and Subsiding Coastal Areas* (eds Millman, J. D. & Haq, B. U.) 11–45 (Kluwer Academic, Norwell, Massachusetts, 1996).
- Wigley, T. M. L., Jones, P. D. & Raper, S. C. B. The observed global warming record: What does it tell us? *Proc. Natl Acad. Sci.* **94**, 8314–8320 (1997).
- Forest, C. E., Allen, M. R., Stone, P. H. & Sokolov, A. P. Constraining uncertainties in climate models using climate change detection techniques. *Geophys. Res. Lett.* **27**, 569–572 (2000).
- Meehl, G. A., Boer, G. J., Covey, C., Latif, M. & Stouffer, R. J. The Coupled Model Intercomparison project (CMIP). *Bull. Am. Meteorol. Soc.* (in the press).
- Santer, B. D. *et al.* A search for human influences on the thermal structure of the atmosphere. *Nature* **382**, 39–46 (1996).
- Hegerl, G. C. *et al.* Detecting greenhouse gas-induced climate change with an optimal fingerprint method. *J. Clim.* **9**, 2281–2306 (1996).
- Hegerl, G. *et al.* On multi-fingerprint detection and attribution of greenhouse gas and aerosol forced climate change. *Clim. Dyn.* **13**, 613–634 (1997).
- North, G. R. & Stevens, M. J. Detecting climate signals in the surface temperature record. *J. Clim.* **11**, 563–577 (1998).
- Tett, S. F. B., Stott, P. A., Allen, M. R., Ingram, W. J. & Mitchell, J. F. B. Causes of twentieth century temperature change near the Earth's surface. *Nature* **399**, 569–572 (1999).
- Johns, T. C. The Second Hadley Centre coupled ocean-atmosphere GCM: model description, spin-up and validation. *Clim. Dyn.* **13**, 103–134 (1997).
- Voss, R., Sausen, R. & Cubasch, U. Periodically synchronously coupled integrations with the atmosphere-ocean general circulation model ECHAM3/LSG. *Clim. Dyn.* **14**, 249–266 (1998).
- Röckner, E., Bengtsson, L., Feichter, J., Lelieveld, J. & Rodhe, H. Transient climate change simulations with a coupled atmosphere-ocean gcm including the tropospheric sulfur cycle. *J. Clim.* **12**, 3004–3032 (1999).
- Knutson, T. R., Delworth, T. L., Dixon, K. W. & Stouffer, R. J. Model assessment of regional surface temperature trends (1949–1997). *J. Geophys. Res.* **104**, 30981–30996 (1999).
- Mitchell, J. F. B., Johns, T. C., Gregory, J. M. & Tett, S. F. B. Climate response to increasing levels of greenhouse gases and sulphate aerosols. *Nature* **376**, 501–504 (1995).
- Sokolov, A. P. & Stone, P. H. A flexible climate model for use in integrated assessments. *Clim. Dyn.* **14**, 291–303 (1998).
- Allen, M. R. Do-it-yourself climate prediction. *Nature* **401**, 642 (1999).
- Hasselmann, K. On multifingerprint detection and attribution of anthropogenic climate change. *Clim. Dyn.* **13**, 601–611 (1997).
- Allen, M. R. & Tett, S. F. B. Checking internal consistency in optimal fingerprinting. *Clim. Dyn.* **15**, 419–434 (1999).
- Stott, P. A. Attribution of twentieth century climate change to natural and anthropogenic causes. *Clim. Dyn.* (in the press).
- Wood, R. A., Keen, A. B., Mitchell, J. F. B. & Gregory, J. M. Changing spatial structure of the thermohaline circulation in response to atmospheric CO₂ forcing in a climate model. *Nature* **399**, 572–575 (1997).
- Mitchell, J. F. B. & Johns, T. C. On modification of global warming by sulphate aerosols. *J. Clim.* **10**, 245–266 (1997).
- Corti, S., Molteni, F. & Palmer, T. N. Signature of recent climate change in frequencies of natural atmospheric circulation regimes. *Nature* **398**, 799–802 (1999).
- Delworth, T. L. & Knutson, T. R. Simulation of early 20th century global warming. *Science* **287**, 2246–2250 (2000).
- Cubasch, U., Voss, R., Hegerl, G. C., Waszkewitz, J. & Crowley, T. J. Simulation of the influence of solar radiation variations on the global climate with an ocean-atmosphere general circulation model. *Clim. Dyn.* **13**, 757–767 (1997).

- Cox, P. M., Betts, R. A., Jones, C. S., Spall, S. A. & Totterdell, I. J. Acceleration of global warming due to carbon-cycle feedbacks in a 3D coupled model. *Nature* (submitted).
- Intergovernmental Panel on Climate Change *Special Report on Emissions Scenarios* (eds Nakicenović, N. & Swart, R.) (Cambridge Univ. Press, 2000).
- Smith, S. J., Wigley, T. M. L., Nakicenovic, N. & Raper, S. C. B. Climate implications of greenhouse gas emissions scenarios. *Technol. Forecast. Social Change* (in the press).
- Stott, P. A. & Tett, S. F. B. Scale-dependent detection of climate change. *J. Clim.* **11**, 3282–3294 (1998).
- Ripley, B. D. & Thompson, M. Regression techniques for the detection of analytical bias. *Analyst* **112**, 377–383 (1987).
- van Huffel, S. & Vanderwaal, J. *The Total Least Squares Problem: Computational Aspects and Analysis* (Society for Industrial & Applied Mathematics, Philadelphia, 1991).

Acknowledgements

We thank T. Barnett, C. Forest, N. Gillett, K. Hasselmann, G. Hegerl, W. Ingram, G. Jones, S. Raper, B. Ripley, S. Smith, A. Sokolov, P. Stone, S. Tett, I. Tracey and A. Weaver for suggestions. This work was supported by the UK Natural Environment Research Council (M.R.A.); The UK Department of Environment, Transport and Regions (P.A.S.); the UK Meteorological Office's Research and Development Programme (J.F.B.T.); the European Commission (R.S.); and the US National Oceanic and Atmospheric Administration (T.L.D.); with additional support from the US Department of Energy and the British Council.

Correspondence and requests for materials should be addressed to M.R.A. (e-mail: m.r.allen@rl.ac.uk).

Pressure-induced changes in the compression mechanism of aluminous perovskite in the Earth's mantle

John P. Brodholt

Department of Geological Sciences, University College London, Gower Street, London WC1E 6BT, UK

Although aluminium is the fifth most abundant element in the Earth's mantle, its effect on the physical properties of perovskite, the main mineral phase in the lower mantle, has largely been ignored. It is becoming clear, however, that many properties of MgSiO₃ perovskites are remarkably sensitive to small amounts of aluminium^{1–4}. In particular, perovskite with only 5 wt% Al₂O₃ has a bulk modulus 10% lower than that of the pure magnesian end-member¹². The increased compressibility may be due to the high concentrations of oxygen vacancies required to balance the charge of the aluminium⁵; if so, this would have important consequences for the mantle, as aluminous perovskites could be weaker, have lower seismic velocities and be hosts for water. To test whether oxygen vacancies exist in aluminous perovskites, I have calculated the compressibility of end-member defect-bearing perovskites using *ab initio* methods. The results show that perovskites with oxygen vacancies do have significantly greater compressibilities than those without such vacancies. But the results also suggest that oxygen vacancies become unfavourable at high pressures, in which case only the physical properties of the shallow lower mantle would be affected by aluminium—with the deeper mantle retaining properties similar to those of aluminium-free perovskite.

Two substitution mechanisms have been proposed for the incorporation of Al³⁺ into MgSiO₃ perovskite. The first is the coupled substitution $2\text{Al}^{3+} \rightarrow \text{Al}_{\text{Mg}}' + \text{Al}_{\text{Si}}'$ (ref. 6) where aluminium enters both cation sites (the M site normally occupied by 2+ cations, and the octahedral site normally occupied by silicon) and does not require the creation of vacancies for charge balance. (Here the ' and ')

superscripts represent a positive or negative net charge on the site, respectively.) The second mechanism is $2\text{Al}^{3+} \rightarrow 2\text{Al}_{\text{Si}}' + \text{V}_{\text{O}}^{\bullet}$ in which aluminium replaces silicon only and oxygen vacancies are required for charge balance. (Here V indicates vacancy.) Atomistic calculations suggest that the coupled substitution mechanism is preferred⁶, and this is supported by XAFS (X-ray absorption fine structure) experiments⁷. Most low-pressure ceramic perovskites, however, incorporate $3+$ ions by the second method³, and this is supported by the presence of excess stishovite in synthetic aluminous perovskites^{8,9}. The greatly enhanced compressibility of aluminous perovskites could throw light on which mechanism is favoured if we knew how the two different mechanisms affect the physical properties of perovskite. In order to evaluate quantitatively how aluminium affects the compressibility of perovskites, I have performed density functional theory (DFT) calculations on two perovskites that represent end-member defect perovskites. The first of these, brownmillerite, is the end-member of a homologous series of oxygen-deficient perovskites that can range from one in six oxygens being vacant to arbitrarily small concentrations of oxygen vacancies¹⁰. The Mg–Al composition brownmillerite has Al^{3+} in the octahedral sites, and Mg^{2+} in the larger cation sites, to give $\text{Mg}_2\text{Al}_2\text{O}_5$ (Fig. 1). The second defect end-member considered is simply a normal orthorhombic MgSiO_3 perovskite, with one Mg and one Si in the unit cell substituted with aluminium; this structure requires no vacancies for charge balance, and the full unit cell is $(\text{Mg}_3\text{Al})(\text{Si}_3\text{Al})\text{O}_{12}$.

Table 1 shows the calculated volumes and cell parameters at 0, 30, 60 and 100 GPa for the two defect perovskites, together with the pure end-member MgSiO_3 perovskite and MgO. As is typical with GGA (generalized gradient approximation) calculations, the zero-pressure volumes are slightly over estimated by 1–2% while the bulk moduli are about 7–8% too low. Figure 2 shows the compressibility of the two defect perovskites together with the magnesium end-member. The compressibility of the $(\text{Mg}_3\text{Al})(\text{Si}_3\text{Al})\text{O}_{12}$ perovskite is similar to the aluminium-free perovskite, and the bulk modulus (K)

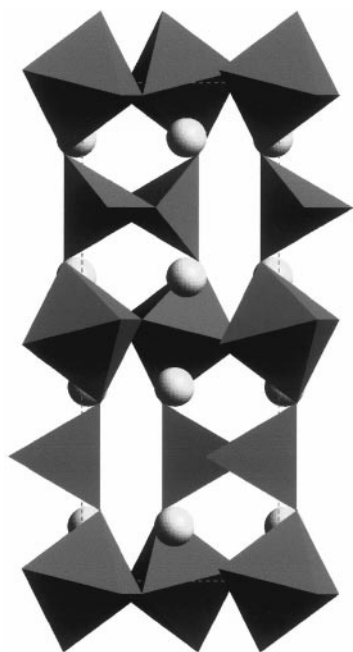
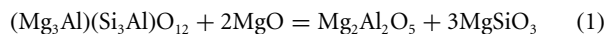


Figure 1 The relaxed zero-pressure brownmillerite $\text{Mg}_2\text{Al}_2\text{SiO}_5$ structure. The balls represent Mg^{2+} , and the tetrahedra and octahedra represent Al^{3+} with respectively four- and six-fold oxygen coordinations. This structure can be made by the removal of one in six oxygens from the normal perovskite structure (not shown), which results in planes of octahedra relaxing to form rows of corner-shared tetrahedra.

is only about 6% lower. As the concentration of Al^{3+} is much greater in this perovskite than in the experimental perovskite, this type of charge-balancing mechanism is not consistent with the experimentally observed increased compressibility. In contrast, the oxygen-vacancy brownmillerite end-member is substantially more compressible, with a K of 155 GPa. A comparison of the theoretical results with the experimental results may be made if we assume that the increase in volume on dissolving alumina into perovskite is linear with composition. (Although this assumption is unlikely to be accurate, it is the best that can be done at the moment.) Adding 5% Al_2O_3 into MgSiO_3 perovskite in the form of the oxygen vacancy brownmillerite structure would give, therefore, a 0 GPa volume of $163.44 \text{ \AA}^3$. This is 1.006 times greater than the pure MgSiO_3 perovskite, and compares well with the experimental increase² of 1.005. Similarly, the bulk modulus would be 225 GPa compared to 239 GPa for the alumina-free perovskite. Although this difference of 6% is not quite as much as the 10% found by Zhang and Weidner², it is certainly more reasonable than the coupled substitution mechanism and indicates the existence of a large percentage (up to half the Al^{3+} concentration) of oxygen vacancies in their experimental samples.

Atomistic calculations have shown that the oxygen defect mechanism is only energetically favourable at low pressures; at lower-mantle pressures, the self-charge-balancing mechanism is more favourable⁶. These simpler calculations ignored the role of MgO, and rather than considering the Al_2O_3 – MgSiO_3 join, it is necessary⁵ to consider the $\text{Mg}_2\text{Al}_2\text{O}_5$ – MgSiO_3 join. This more realistic composition will tend to increase the concentration of oxygen vacancies in the perovskite. In order to evaluate the comparative stability of the two end-member defect perovskites, I have calculated the enthalpy change (ΔH) of the reaction



at zero temperature and pressures from 0 to 100 GPa. The results are shown in Fig. 3. As with the atomistic calculations that ignore MgO, the oxygen vacancy mechanism is stable at low pressures and the self-charge-balancing mechanism is stable at high pressures. The difference, however, is that MgO buffering has increased the stability of the oxygen vacancy mechanism so that it can occur in the lower mantle. The right-hand side of equation (1) may be further stabilized by a small enthalpy of solution, and a true comparison must incorporate entropic effects. These, however,

Table 1 Summary of cell parameters and volumes

	<i>P</i> (GPa)	<i>a</i> (Å)	<i>b</i> (Å)	<i>c</i> (Å)	<i>V</i> (Å ³)
MgSiO_3	0	4.79	4.92	6.89	162.42
$K = 239 \text{ GPa}$	30	4.62	4.77	6.66	146.98
$K_{\text{exp}} = 261 \text{ GPa (ref. 11)}$	60	4.50	4.68	6.49	136.48
	100	4.36	4.57	6.32	126.09
$(\text{Mg}_3\text{Al})(\text{Si}_3\text{Al})\text{O}_{12}$	0	4.77	4.92	6.95	163.31
$K = 224 \text{ GPa}$	30	4.59	4.78	6.69	146.86
	60	4.46	4.68	6.51	135.91
	100	4.32	4.59	6.33	125.47
$\text{Mg}_2\text{Al}_2\text{O}_5$	0	5.08	13.20	5.14	345.16
$K = 155 \text{ GPa}$	30	4.91	12.55	4.86	299.48
	60	4.81	12.05	4.71	272.71
	100	4.72	11.57	4.54	247.93
MgO	0	4.27	4.27	4.27	77.90
$K = 147 \text{ GPa}$	30	4.07	4.07	4.07	67.21
$K_{\text{exp}} = 160 \text{ GPa (ref. 13)}$	60	3.93	3.93	3.93	60.90
	100	3.91	3.91	3.91	55.30

Cell parameters (*a*, *b* and *c*) and volumes (*V*) for the two defect end-member perovskites together with MgSiO_3 perovskite and MgO at pressures of 0, 30, 60 and 100 GPa. The bulk moduli, K , are from a fit to a Birch–Murnaghan equation of state with $dK/dP = 4$. Calculations were performed using the commercially available Molecular Simulations CASTEP density functional (DFT) plane wave code¹⁴. The supplied Vanderbilt ultrasoft pseudopotentials were used with a plane-wave cut-off of 380 eV. Tests with a 900-eV cut-off gave negligible differences for volumes, and the reaction enthalpy (see text) agreed to within 0.04 eV with the 380-eV cut-off calculation; this is insignificant for this work. A Monkhorst–Pack *k*-point grid of $0.1 \text{ \AA}^{-1}$ was used for all structures. Increasing the *k*-point grid also produced negligible differences in the equations of state and reaction enthalpies.

are unlikely to overcome the ΔH of 1.8 eV at the base of the lower mantle.

If, as suggested here, aluminium is incorporated by an oxygen vacancy mechanism in the shallow lower mantle and by a vacancy-free mechanism in the deeper lower mantle, we would expect to see a variation of physical properties with depth. Assuming an isochemical lower mantle, the shallow part of the lower mantle will be more compressible and have a larger thermal expansion than the lower mantle. In addition, oxygen vacancies could enhance oxygen diffusion, and as a result, the creep mechanism may also be different and the shallower part of the lower mantle could be significantly weaker than the deeper part. The shallower part of the lower mantle would also be able to incorporate more water in reactions such as⁵

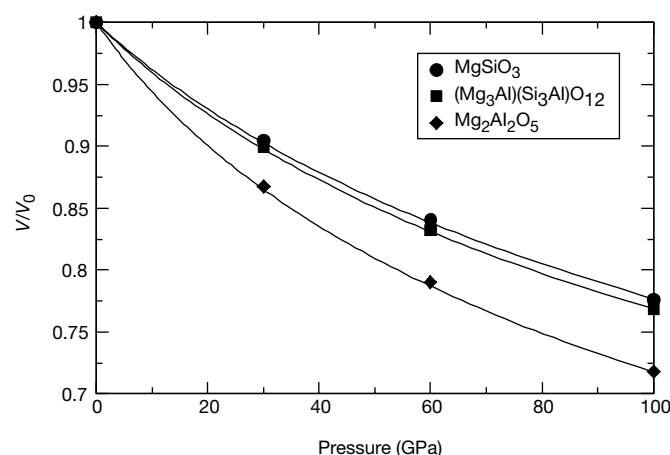


Figure 2 The compressibility (V/V_0) of the three end-member perovskites studied here. (Here V indicates volume, and V_0 indicates volume at 0 GPa.) The perovskite with Al^{3+} ions in both the M site and the octahedral site (squares) has a compressibility that is only slightly greater than the MgSiO_3 perovskite (circles). Given that the Al^{3+} concentration in this perovskite is significantly greater than in the experimental perovskites of ref. 2, it seems unlikely that this sort of charge-balancing mechanism could be responsible for the experimentally observed increased compressibility. The $\text{Mg}_2\text{Al}_2\text{O}_5$ brownmillerite structure (diamonds), on the other hand, is significantly more compressible, and supports the explanation that oxygen vacancies are responsible for the strong increase in compressibility in the experimental alumina-rich perovskites⁵.

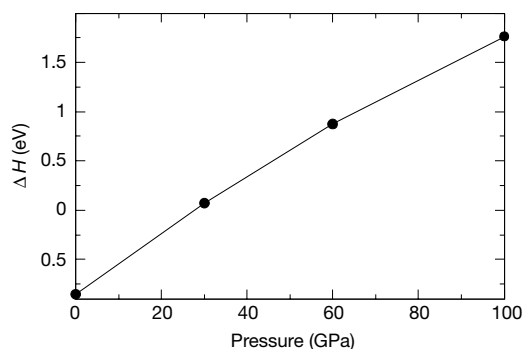


Figure 3 The enthalpy difference between the oxygen vacancy perovskite (brownmillerite) and the self-charge-balanced perovskite as a function of pressure. This is ΔH for the reaction shown in equation (1) in the text. At pressures below 30 GPa, the oxygen vacancy mechanism is favoured, while at higher pressures the coupled substitution is favoured. These calculations are at 0 K; to calculate the pressure of the transition at the temperatures of the Earth's mantle would require a full evaluation of the entropic effects. Nevertheless, these athermal results suggest that there may be a transition between more compressible oxygen-vacancy-rich perovskite to normal, less-compressible perovskite in the Earth's lower mantle.

$\text{V}_\text{O}'' + \text{O}^{2-} + \text{H}_2\text{O} = 2\text{OH}^-$. The results here give a transition pressure from oxygen-vacancy to vacancy-free perovskite at around 30 GPa; the high-temperature transition pressure would be higher if entropy favours the oxygen vacancy mechanism. There is some indication that the transition pressure is higher from experiments on perovskites with 20% of the cations as aluminium¹¹. At 55 GPa, the total cation sum is slightly higher than the ideal stoichiometric value of two based on three oxygens, which indicates a small oxygen deficiency. At 60 and 70 GPa, however, the total cation sum is the stoichiometric proportion of two, which suggests that aluminium is incorporated in both the octahedral and dodecahedral sites. These results are only just about significant, as the reported analytical uncertainties are about 1% for the oxides. Nevertheless, if taken at face value, it is tempting to suggest that the transition to vacancy free perovskites would occur at around 55–60 GPa.

To understand fully the implications for the mantle, we also need to consider the effect of iron, especially ferric iron as it has been shown that the substitutions of Fe^{3+} and Al^{3+} in perovskites are related^{3,4,6}. The nature of this dependence, however, has not yet been clarified. Although ferric iron in the large dodecahedral sites could balance the charge of the Al^{3+} on the octahedral site and thereby remove the requirement for oxygen vacancies, this would require the same number of Fe^{3+} ions as Al^{3+} . But Mössbauer results on experimental perovskites with aluminium and iron report only two-thirds of the iron that is needed in the 3+ state to balance the charge of all the Al^{3+} (ref. 3). In addition, more recent Mössbauer results on aluminium-free perovskites suggest that at low oxygen fugacity ferric iron prefers the smaller octahedral site¹² and therefore would not balance the charge of the aluminium. Further experiments are needed to elucidate fully the role of aluminium and ferric iron in mantle perovskites, but nevertheless, this work shows that mantle perovskites have a complex crystal chemistry that may have important consequences for models of the Earth's mantle. □

Received 13 January; accepted 24 August 2000.

- Xu, Y. S., McCammon, C. & Poe, B. T. The effect of alumina on the electrical conductivity of silicate perovskite. *Science* **282**, 922–924 (1998).
- Zhang, J. & Weidner, D. J. Thermal equation of state of aluminium-enriched silicate perovskite. *Science* **284**, 782–784 (1999).
- McCammon, C. A. Perovskite as a possible sink for ferric iron in the lower mantle. *Nature* **387**, 694–696 (1997).
- Wood, B. J. & Rubie, D. C. The effect of alumina on phase transformations at the 660-kilometer discontinuity from Fe–Mg partitioning experiments. *Science* **273**, 1522–1524 (1996).
- Navrotsky, A. A lesson from ceramics. *Science* **284**, 1788–1789 (1999).
- Richmond, N. C. & Brodholt, J. P. Calculated role of aluminium in the incorporation of ferric iron into magnesium silicate perovskite. *Am. Miner.* **83**, 947–951 (1998).
- Andrault, D., Neuville, D. R., Flank, A. M. & Wang, Y. Cation sites in Al-rich MgSiO_3 perovskites. *Am. Miner.* **83**, 1045–1053 (1998).
- Ito, E., Kubo, A., Katsura, M., Akaogi, M. & Fujita, T. High-pressure transformation of pyrope ($\text{Mg}_3\text{Al}_2\text{Si}_3\text{O}_{12}$) in a sintered diamond cubic anvil assembly. *Geophys. Res. Lett.* **25**, 821–824 (1998).
- McCammon, C., Langenhorst, F. & Seifert, F. How well do we know the properties of lower mantle perovskites? *Eos* **80**, F742 (1999).
- Smyth, D. M., Chang, E. K. & Liu, D. H. Travels through perovskite space. *Phase Transit.* **58**, 57–73 (1996).
- Kesson, S. E., Gerald, J. D. F., Shelley, J. M. G. & Withers, R. L. Phase-relations, structure and crystal-chemistry of some aluminous silicate perovskites. *Earth. Planet. Sci. Lett.* **134**, 187–201 (1995).
- McCammon, C. A. The crystal chemistry of ferric iron in $\text{Fe}_{0.05}\text{Mg}_{0.95}\text{SiO}_3$ perovskite as determined by Mössbauer spectroscopy in the temperature range 80–293 K. *Phys. Chem. Miner.* **25**, 292–300 (1998).
- Fei, Y. Effects of temperature and composition on the bulk-modulus of $(\text{Mg},\text{Fe})\text{O}$. *Am. Miner.* **84**, 272–276 (1999).
- CASTEP version 3.9 (Molecular Simulations Inc., San Diego, 1999).

Acknowledgements

I thank C. McCammon for comments on this manuscript, N. Ross for her help, and D. Dobson, A. Oganov, L. Vocablo, I. Wood and D. Price for discussions. I also thank the Royal Society for a University Research Fellowship.

Correspondence and requests for materials should be addressed to the author (e-mail: j.brodholt@ucl.ac.uk).

A marine microbial consortium apparently mediating anaerobic oxidation of methane

Antje Boetius*, Katrin Ravensschlag*, Carsten J. Schubert*, Dirk Rickert†, Friedrich Widdel*, Armin Gieseke*, Rudolf Amann*, Bo Barker Jørgensen*, Ursula Witte* & Olaf Pfannkuche†

* Max Planck Institute for Marine Microbiology, 28359 Bremen, Germany

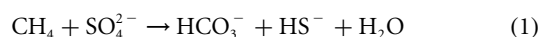
† GEOMAR Research Center for Marine Geosciences, 24148 Kiel, Germany

A large fraction of globally produced methane is converted to CO₂ by anaerobic oxidation in marine sediments¹. Strong geochemical evidence for net methane consumption in anoxic sediments is based on methane profiles², radiotracer experiments³ and stable carbon isotope data⁴. But the elusive microorganisms mediating this reaction have not yet been isolated, and the pathway of anaerobic oxidation of methane is insufficiently understood. Recent data suggest that certain archaea reverse the process of methanogenesis by interaction with sulphate-reducing bacteria^{5–7}. Here we provide microscopic evidence for a structured consortium of archaea and sulphate-reducing bacteria, which we identified by fluorescence *in situ* hybridization using specific 16S rRNA-targeted oligonucleotide probes. In this example of a structured archaeal-bacterial symbiosis, the archaea grow in dense aggregates of about 100 cells and are surrounded by sulphate-reducing bacteria. These aggregates were abundant in gas-hydrate-rich sediments with extremely high rates of methane-based sulphate reduction, and apparently mediate anaerobic oxidation of methane.

At the Cascadia convergent margin off the coast of Oregon, discrete methane hydrate layers are exposed at the sea floor, at a water depth of 600–800 m that corresponds to the hydrate stability limit⁸. These hydrate layers are formed from gaseous methane

that continuously ascends along faults generated by accretionary tectonics. The crest of the southern Hydrate Ridge (44° 34' N, 125° 09' W, 780 m water depth) is populated by large communities of clams of the genus *Calyptogena*, and by thick bacterial mats of the sulphide-oxidizing *Beggiatoa*, both of which indicate areas of active gas seeping⁹. Undisturbed sediment cores with *Beggiatoa* mats were obtained using a video-guided multiple corer during RV SONNE Cruise SO143-2 in August 1999 (ref. 10). These samples often released gas bubbles due to decompression during recovery.

Sulphate reduction rates (SRRs) were extremely high in sediments covered by *Beggiatoa* mats, reaching more than 5 μmol cm⁻³ d⁻¹ in the surface sediments (Fig. 1). Integrated over the upper 15 cm, the resulting SRR is 140 mmol m⁻² d⁻¹; this is, to our knowledge, the highest value ever measured in cold marine sediments. At a nearby reference station without gas hydrates and vent colonization, SRRs were below the detection limit (< 1 nmol cm⁻³ d⁻¹). Thus, at Hydrate Ridge, sulphate reduction is clearly fuelled by high methane fluxes from below, while organic deposition from surface waters is not a significant substrate source for sulphate-reducing bacteria (SRB). A similar phenomenon was observed at gas seeps in the Gulf of Mexico, where 600-fold higher SRRs were measured at methane seeps (up to 2.5 μmol cm⁻³ d⁻¹, calculated from sulphate concentration profiles) compared to reference stations¹¹. The restriction of such high SRRs to sediments rich in methane is evidence for a direct link between the processes of methane and sulphate turnover. Sulphate has been proposed to be the terminal electron acceptor in the zone of anaerobic oxidation of methane³, according to:



Assuming this stoichiometry, the turnover of methane can exceed 5 mM d⁻¹ in the sediments of the Hydrate Ridge, where a dissolved methane concentration of 80 mM is reached above decomposing gas hydrates at *in situ* temperature (4 °C) and hydrostatic pressure (8 MPa). As one product of anaerobic oxidation of methane, sulphide accumulates to concentrations almost equivalent to those of sulphate consumed (Fig. 1). Intense sulphide production

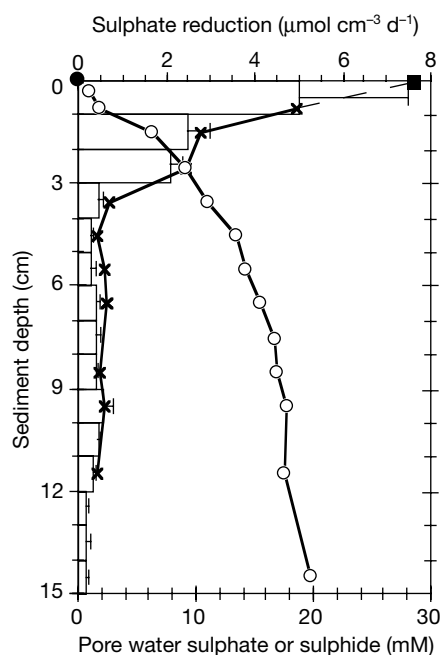


Figure 1 Depth profiles of various quantities from methane-rich sediments. Crosses, porewater sulphate concentrations (bottom axis); open circles, porewater dissolved sulphide concentrations (bottom axis); filled circle and filled square, concentrations in overlying bottom water of sulphate and sulphide, respectively. Columns with error bars

show average values of sulphate reduction rate (SRR). The multiple corer SO143/173-1B-TVMC was positioned on a *Beggiatoa* mat¹⁰. One core was taken for porewater chemistry, and three replicate subcores were obtained for SRR. Error bars indicate the standard deviation between the subcores.

explains the occurrence of sulphide-based *Beggiatoa*/*Calymmatobacter* communities at the Hydrate Ridge. The methane-derived bicarbonate precipitates as calcium carbonate, and forms large structures exposed at the crest of the Hydrate Ridge⁸.

In Hydrate Ridge sediments, the archaeal isoprenoids crocetane and pentamethylcosane were found to be highly depleted in ¹³C ($\delta^{13}\text{C} = -124\text{‰}$ vs PDB)¹². Additionally, in *Beggiatoa*-covered sediments of 0–10 cm depth, we found archaeol and *sn*-2-hydroxyarchaeol at high concentrations (8 μg per g sediment dry mass) and similarly depleted in ¹³C (-114‰ and -133‰ , respectively). These lipids are common in archaea, and are particularly prominent in methanogens¹³. Such highly ¹³C depleted lipid biomarkers are

due to consumption of methane with a $\delta^{13}\text{C}$ of -62‰ to -72‰ (ref. 8) and subsequent fractionation. Very light iso- and anteiso- C_{15} fatty acids (-63‰ and -75‰ , respectively) which occur abundantly in SRB were also detected at high concentrations (10 μg per g sediment dry weight). These values are similar to, or even lower than, values reported from methane-rich environments, such as a gas-hydrate-bearing seep⁷, a Miocene limestone from an ancient vent system¹⁴, and an active mud volcano¹⁵.

In the *Beggiatoa*-covered sediments of the Hydrate Ridge, abundant cell aggregates were detected by fluorescence *in situ* hybridization (FISH) specific for the domain Archaea¹⁶. These cell aggregates were not found at the reference station without methane seepage. The archaeal cells in the aggregates were detected with probe EelMS932 targeting clone sequences which were retrieved from a similar methane-rich environment (Eel River basin, California)⁷ and which are phylogenetically affiliated with the order Methanosarcinales. The aggregated archaea were poorly stained with 4',6'-diamidino-2-phenylindole (DAPI; Fig. 2a, c, e), and were recognized as such only by the probe signal (Fig. 2b, d, f). Specific FISH analysis of the outer layer of DAPI-stained cells revealed that these are members of the domain Bacteria¹⁷, and belong to the SRB of the delta-proteobacteria. The SRB surrounding the archaeal aggregates were targeted with probe DSS658 (Fig. 2g) specific for the branch *Desulfosarcina*/*Desulfococcus*¹⁸, and with probe DSS225 which is highly specific for a new subgroup of that branch¹⁹. The closest cultivated relative of this subgroup is *Desulfosarcina variabilis* with up to 91.2% 16S rDNA sequence similarity.

An average archaea/SRB consortium consisted of an inner sphere of $2.3 \pm 1.3 \mu\text{m}$ diameter containing about 100 coccoid archaeal cells, each $0.5 \mu\text{m}$ in diameter. These were partially or fully surrounded by about 200 cells of SRB ($0.3\text{--}0.5 \mu\text{m}$ in diameter), which formed an outer shell of mostly 1–2 cell layers. The size spectrum of 100 archaea/SRB consortia ranged from 1 to 11 μm in diameter with an average of $3.2 \pm 1.5 \mu\text{m}$. The smallest aggregates consisted of only 1–3 archaeal cells and 1–3 cells of SRB, and may represent early stages of the consortium development, while the largest contained about 10,000 cells. The consortia were highly abundant in surface sediments at sulphide concentrations $< 10 \text{ mM}$ (Fig. 1), with a maximum of 7×10^7 aggregates cm^{-3} at 1–2 cm depth (Fig. 3). The average number in the upper 5 cm was 3×10^7 aggregates cm^{-3} , which is equivalent to about 3×10^9 archaeal cells cm^{-3} and 6×10^9 cells of SRB cm^{-3} . Hence, the consortia comprised 94% of all archaea detected with domain-specific probe ARCH915¹⁶ and 96%

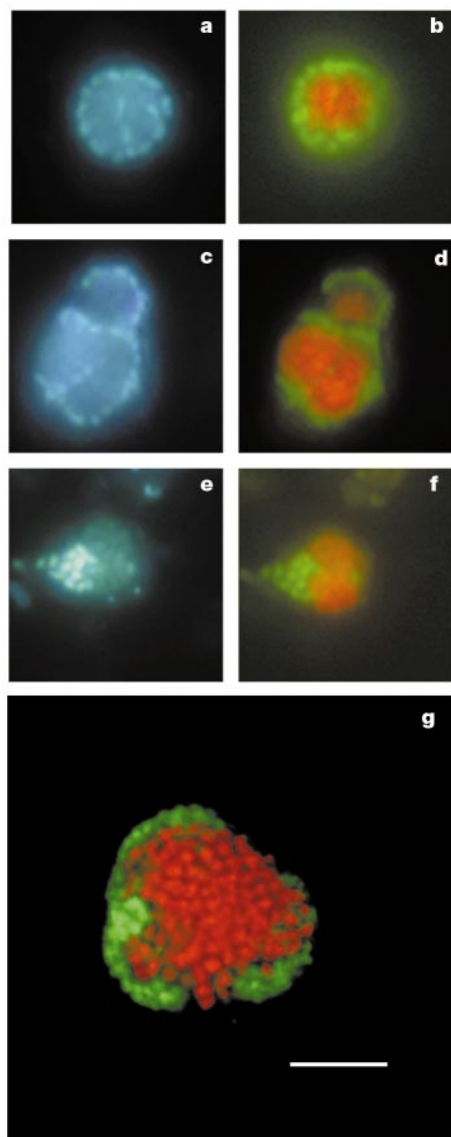


Figure 2 *In situ* identification of archaea/SRB aggregates with fluorescently labelled rRNA-targeted oligonucleotide probes. The archaea are shown in red, and the SRB in green. The aggregates were visualized using filter sets specific for DAPI, CY3 and FLUOS for identical microscopic fields. Single xy images of the same optical section were combined. **a/b, c/d, e/f**, Epifluorescence micrographs of different aggregates stained with DAPI (left panel) and hybridized with the CY3-labelled probe EelMS932 (5'-AGCT CCACCGTTGTAGT-3') and the FLUOS-labelled probe DSS658 (5'-TCCACTTCCCTCT CCCAT-3')¹⁹ (right panel). After testing for their specificity at ~60% v/v formamide, both probes were hybridized at 40% formamide to give optimal brightness. Scale bar, 5 μm . **g**, Confocal laser scanning micrograph of the hybridization with the CY3-labelled probe EelMS932 (archaea) and the FLUOS-labelled probe DSS658 (SRB). Details of probes are available, see Supplementary Information. Scale bar, 5 μm .

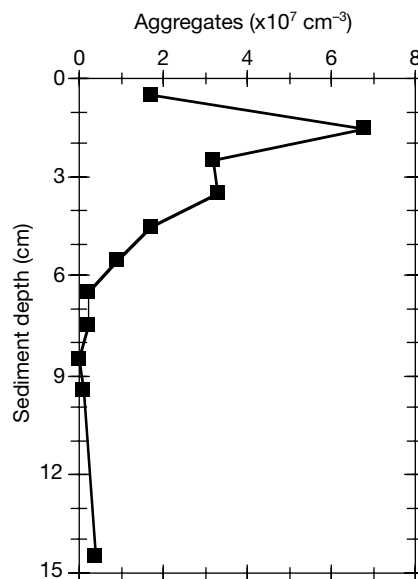


Figure 3 Abundance of archaea/SRB consortia in a sediment core from a *Beggiatoa* mat.

of all SRB detected with eight different genus-specific probes^{18,20,21}. From the abundance and biovolume of cells in the consortia, a biomass of SRB of 0.12 mg cell dry mass per cm³ is calculated. Hence, the SRR of 5 $\mu\text{mol cm}^{-3} \text{ d}^{-1}$ would yield a specific rate of 42 $\mu\text{mol per mg cell dry mass per day}$. This value is within the range of specific rates in cultures of SRB grown under optimal conditions in the laboratory (15–430 $\mu\text{mol per mg cell dry mass per day}$)²².

To identify other organisms potentially consuming methane in the Hydrate Ridge sediments, we developed specific probes for FISH of other archaea and methylophilic bacteria. Four different probes for the newly described phylogenetic ANME-1 cluster, suspected of consuming methane anaerobically in Eel River basin sediments⁷, did not hybridize with the aggregated archaea. Hybridization of non-aggregated archaea with ANME-1 probes was far below 1% of the total DAPI cell counts at lowest stringency (0% formamide). Furthermore, no other methanogenic archaea of the orders Methanosarcinales and Methanobacteriales were detected by FISH using specific 16S rRNA probes tested with appropriate reference organisms. FISH counts of aerobic methylophilic of the alpha- and gamma-proteobacteria were also below detection limits at all stations. Thus, aerobic oxidation of methane by relatives of known methanotrophs does not appear to be an important process in the methane-rich sediments of the Hydrate ridge. Rather, the abundant, strongly ¹³C depleted, consortia of methanogenic archaea and SRB are mediating the anaerobic oxidation of methane. We assume that this process is a reversal of methane formation, involving methanogens and a sulphate reducing partner which effectively scavenges intermediates such as H₂ or acetate^{5,6,26}.

So far, only few examples of prokaryotic symbioses based on metabolic interaction in direct cell contact have been identified; for example, the microbial consortium "Chlorochromatium"²³ and the cluster of the nitrifying bacteria *Nitrosomonas* and *Nitrobacter*²⁴. The advantage of the archaea/SRB consortium compared with free-living cells would be a highly efficient transfer of intermediates by molecular diffusion²⁵. Members of the Methanosarcinales, as observed in the consortia, encompass metabolically diverse methanogens. They may form methane not only from CO₂ and H₂, but also from simple methyl-group-containing compounds such as acetate, methanol, methylamines and methyl sulphide (but not from formate which is used by different methanogens). Members of the *Desulfosarcina/Desulfococcus* branch include nutritionally versatile SRB that oxidize organic compounds (including acetate) completely to CO₂, and several species can grow autotrophically with CO₂, H₂ and sulphate.

The low $\delta^{13}\text{C}$ signature of SRB lipids is best explained if reverse methanogenesis leads to an organic intermediate that serves not only as electron donor for sulphate reduction, but also as a cellular carbon source. A hypothesized bimolecular reaction of methane, in which both carbon atoms of acetate are derived from methane coupled to simultaneous H₂ formation²⁶, would offer the most favourable explanation for a strong ¹³C-depletion in SRB. But with any organic intermediate, the cellular carbon of SRB would still be partly derived from CO₂ via some biosynthetic reactions. However, at the high methane oxidation rate estimated from our SRR measurements, the densely packed consortia may maintain a methane-derived isotopically light CO₂/HCO₃⁻ pool (equation (1)) that is not fully equilibrated with the surrounding heavier porewater pool. Furthermore, autotrophic growth of SRB using the carbon monoxide dehydrogenase pathway (present in members of the *Desulfosarcina* branch) for CO₂ fixation is associated with pronounced ¹³C discrimination²⁷. Hence, reverse methanogenesis via CO₂ and H₂ may remain relevant as a model.

The free energy change by anaerobic oxidation of methane (equation (1)) at the pressure (about 8 MPa) in the zone with the highest numbers of aggregates is $\Delta G = -40 \text{ kJ per mol methane}$ ($\Delta G = -30 \text{ kJ}$ at the methane pressure of 0.1 MPa prevailing after sampling). The intermediate concentrations required to keep both

reverse methanogenesis and sulphate reduction energetically feasible are 10⁻¹⁰ to 10⁻⁹ M (0.01 to 0.1 Pa) for hydrogen and 3 × 10⁻¹² to 3 × 10⁻⁸ M for acetate. This requirement cannot easily be reconciled with the observed high SRR. Due to the lack of reliable half-saturation constants (K_M values) of SRB and mass transfer rates within densely packed consortia, the kinetics of this system, involving extremely low intermediate concentrations, remain uncertain. The maintenance of a microenvironment strongly depleted in hydrogen, acetate or other possible intermediates is probably a prerequisite for the anaerobic oxidation of methane via reversed methanogenesis in the consortia. □

Methods

Sulphate reduction rates

Sediment cores were immediately transferred to a cold room (4 °C), and ³⁵SO₄²⁻ was injected horizontally into the intact sediment cores at 1-cm depth intervals. The cores were incubated for 24 h at *in situ* temperature before the reaction was stopped by mixing the sediments with 20% zinc acetate. The samples were stored frozen until the single-step acidic distillation of the Cr(II)-reduced sulphur compounds was performed as described in ref. 28. SRR was calculated from the ratio of radioactive sulphide to the total radioactive sulphate added. Measurements of porewater sulphate and sulphide were performed as described previously⁸.

Lipid analysis

Freeze-dried and gently ground sub-samples were extracted by successive sonication and centrifugation in methanol, methanol:methylene chloride (1:1) and methylene chloride. After saponification (6% KOH), the neutral fraction was extracted with hexane, and derivatized with bis-(trimethylsilyl)-trifluoroacetamide (BSTFA; Sigma) before injection onto a HP5 chromatographic column (30 m length, 0.32 mm i.d., 0.17 μm film thickness). Column temperature was programmed from 90 to 180 °C at a rate of 10 °C min⁻¹ and then at a rate of 6 °C min⁻¹ to 320 °C (30 min isothermal). The acid fraction was recovered after adding HCl and fatty acids were transferred to fatty acid methyl esters with BF₃-methanol. Chromatographic conditions were the same as for the neutral fraction. Individual compounds were identified on a Finnigan MAT GCQ ion trap. Stable carbon isotopes were determined under the same chromatographic conditions with a HP6890 gas chromatograph coupled to a Finnigan Delta Plus isotope mass spectrometer. Reported δ values are corrected for the introduction of additional carbon atoms by derivatization with either BSTFA or BF₃-MeOH.

FISH

Sediment cores from methane-rich sites and from a reference site not enriched in methane were sliced into 1-cm sections. Samples were fixed for 2–3 h with 4% formaldehyde, washed twice with PBS (10 mM sodium phosphate; 130 mM NaCl) and finally stored in PBS/EtOH (1:1) at -20 °C. Stored samples were diluted and treated by mild sonication for 20 s with a MS73 probe (Sonopuls HD70, Bandelin) at an amplitude of 42 μm and a power of < 10 W. An aliquot was filtered on 0.2- μm polycarbonate filters (GTP, Millipore). Hybridization and microscopy counts of hybridized and DAPI-stained cells were performed as described previously²⁹. CY3- and carboxyfluorescein- (FLUOS-) labelled oligonucleotides were obtained from Interactiva (Germany).

Calculation of biomass and rates

The biovolume and biomass of SRB were calculated assuming a spherical cell shape with a diameter of 0.5 μm . Consequently, 6 × 10⁹ cells would have a biovolume of 0.39 mm³. With a wet mass/volume ratio of approximately 1 mg mm⁻³ and a conversion coefficient of 0.3 mg dry mass per mg wet mass, the cell dry mass per cm³ of sediment was 0.12 mg.

Calculation of free energy

Free energy changes (ΔG values) were calculated from G_f° data³⁰ via ΔG° values. Calculations were done for the indicated methane pressures, a temperature of 4 °C, a pH of 7.5 (if H⁺ ions are involved), and average concentrations of SO₄²⁻, HCO₃⁻ and HS⁻ of 2 × 10⁻², 1 × 10⁻² and 2 × 10⁻³ M, respectively (as prevailing at the sediment depth with the highest number of aggregates). For SO₄²⁻, HCO₃⁻ and HS⁻ in sea water, activity coefficients of 0.1, 0.5 and 0.5, respectively, were estimated³⁰. The influence of temperature on ΔG° (for example, for equation (1): $\Delta G^\circ_{277\text{K}} = -16.2 \text{ kJ}$ versus $\Delta G^\circ_{298\text{K}} = -16.6 \text{ kJ}$ per mol methane oxidized) was calculated by the integrated Gibbs-Helmholtz equation including enthalpy (ΔH°) values³⁰. H₂ pressures and acetate concentrations were calculated that allow a free energy threshold of approximately -10 kJ per mol methane for each partner involved in the overall reaction (equation (1)).

Received 16 May; accepted 1 August 2000.

1. Reeburgh, W. S. in *Dynamic Environment of the Ocean Floor* (eds Fanning, K. & Manheim, F. T.) 203–217 (Heath, Lexington, Massachusetts, 1982).
2. Martens, C. S. & Berner, R. A. Interstitial water chemistry of Long Island Sound sediments. I. Dissolved gases. *Limnol. Oceanogr.* **22**, 10–25 (1977).

3. Iversen, N. & Jørgensen, B. B. Anaerobic methane oxidation rates at the sulfate-methane transition in marine sediments from Kattegat and Skagerrak (Denmark). *Limnol. Oceanogr.* **30**, 944–955 (1985).
4. Alperin, M. J., Reeburgh, W. S. & Whiticar, M. J. Carbon and hydrogen isotope fractionation resulting from anaerobic methane oxidation. *Glob. Biochem. Cycles* **2**, 279–288 (1988).
5. Hoehler, T. M., Alperin, M. J., Albert, D. B. & Martens, C. S. Field and laboratory studies of methane oxidation in an anoxic marine sediment: Evidence for a methanogenic-sulfate reducer consortium. *Glob. Biogeochem. Cycles* **8**, 451–463 (1994).
6. Hansen, L. B., Finster, K., Fossing, H. & Iversen, N. Anaerobic methane oxidation in sulfate depleted sediments: effects of sulfate and molybdate additions. *Aquat. Microb. Ecol.* **14**, 195–204 (1998).
7. Hinrichs, K. U., Hayes, J. M., Sylva, S. P., Brewer, P. G. & DeLong, E. F. Methane-consuming archaeobacteria in marine sediments. *Nature* **398**, 802–805 (1999).
8. Suess, E. *et al.* Gas hydrate destabilization: enhanced dewatering, benthic material turnover and large methane plumes at the Cascadia convergent margin. *Earth Planet. Sci. Lett.* **170**, 1–5 (1999).
9. Linke, P. *et al.* In situ measurement of fluid flow from cold seeps at active continental margins. *Deep Sea Res.* **41**, 721–739 (1994).
10. Bohrmann, G., Linke, P., Suess, P. & Pfannkuche, O. RV SONNE Cruise Report SO143: TECFLUX-I-1999. *GEOMAR Rep.* **93**, (2000).
11. Aharon, P. & Fu, B. Microbial sulfate reduction rates and sulfur and oxygen isotope fractionations at oil and gas seeps in deepwater Gulf of Mexico. *Geochim. Cosmochim. Acta* **64**, 233–246 (2000).
12. Elvert, M., Suess, E. & Whiticar, M. J. Anaerobic methane oxidation associated with marine gas hydrates: superlight C-isotopes from saturated and unsaturated C₃₀ and C₂₅ irregular isoprenoids. *Naturwissenschaften* **86**, 295–300 (1999).
13. Koga, Y., Morii, H., Akagawa-Matsushita, M. & Ohga, M. Correlation of polar lipid composition with 16S rRNA phylogeny in methanogens. Further analysis of lipid component parts. *Biosci. Biotechnol. Biochem.* **62**, 230–236 (1998).
14. Thiel, V. *et al.* Highly isotopically depleted isoprenoids: molecular markers for ancient methane venting. *Geochim. Cosmochim. Acta* **63**, 3959–3966 (1999).
15. Pancost, R. D. *et al.* Biomarker evidence for widespread anaerobic methane oxidation in Mediterranean sediments by a consortium of methanogenic archaea and bacteria. *Appl. Environ. Microbiol.* **66**, 1126–1132 (2000).
16. Amann, R. L., Krumholz, L. & Stahl, D. A. Fluorescent-oligonucleotide probing of whole cells for determinative phylogenetic and environmental studies in microbiology. *J. Bacteriol.* **172**, 762–770 (1990).
17. Amann, R. L. *et al.* Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Appl. Environ. Microbiol.* **56**, 1919–1925 (1990).
18. Manz, W., Eisenbrecher, M., Neu, T. R. & Szwed, U. Abundance and spatial organization of Gram-negative sulfate-reducing bacteria in activated sludge investigated by in situ probing with specific 16S rRNA targeted oligonucleotides. *FEMS Microbiol. Ecol.* **25**, 43–61 (1998).
19. Ravensschlag, K., Sahm, K., Knoblauch, C., Jørgensen, B. B. & Amann, R. Community structure, cellular rRNA content and activity of sulfate-reducing bacteria in marine Arctic sediments. *Appl. Environ. Microbiol.* **66**, 3590–3600 (2000).
20. Sahm, K., Knoblauch, C. & Amann, R. Phylogenetic affiliation and quantification of psychrophilic sulfate-reducing isolates in marine arctic sediments. *Appl. Environ. Microbiol.* **65**, 3976–3981 (1999).
21. Devereux, R., Kane, M. D., Winfrey, J. & Stahl, D. A. Genus- and group-specific hybridization probes for determinative and environmental studies of sulfate-reducing bacteria. *Syst. Appl. Microbiol.* **15**, 601–609 (1992).
22. Rabus, R., Hansen, T. A. & Widdel, F. in *The Prokaryotes: An Evolving Electronic Resource for the Microbiological Community* (eds Dworkin, M., Falkow, S., Rosenberg, E., Schleifer, K.-H. & Stackebrandt, E.) (Springer, New York, 2000).
23. Fröstl, J. M. & Overmann, J. Physiology and tactic response of the phototrophic consortium “*Chlorochromatium aggregatum*”. *Arch. Microbiol.* **169**, 129–135 (1998).
24. Mobarry, B., Wagner, M., Urbain, V., Rittmann, B. E. & Stahl, D. A. Phylogenetic probes for analyzing abundance and spatial organization of nitrifying bacteria. *Appl. Environ. Microb.* **62**, 2156–2162 (1996).
25. Boone, D. R., Johnson, R. L. & Liu, Y. Diffusion of the interspecies electron carriers H₂ and formate in methanogenic ecosystems and its implications in the measurement of K_m and H₂ or formate uptake. *Appl. Environ. Microbiol.* **55**, 1735–1741 (1989).
26. Valentine, D. L. & Reeburgh, W. S. New perspectives on anaerobic methane oxidation. *Environ. Microb.* **2** (in the press).
27. Preuß, A., Schauder, R. & Fuchs, G. Carbon isotope fractionation by autotrophic bacteria with three different CO₂ fixation pathways. *Z. Naturforsch. C* **44**, 397–402 (1989).
28. Fossing, H. & Jørgensen, B. B. Measurement of bacterial sulfate reduction in sediments: Evaluation of a single-step chromium reduction method. *Biogeochemistry* **8**, 205–222 (1989).
29. Snaird, J., Amann, R., Huber, I., Ludwig, W. & Schleifer, K. H. Phylogenetic analysis and in situ identification of bacteria in activated sludge. *Appl. Environ. Microbiol.* **63**, 2884–2896 (1997).
30. Stumm, W. & Morgan, J. J. *Aquatic Chemistry* 2nd edn (Wiley, New York, 1981).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

The expedition SO143 was performed as part of the program TECFLUX, which addresses geophysical, biogeochemical and hydrographic processes associated with fluid venting from the Cascadia continental margin. The field work of TECFLUX'99 was jointly planned, coordinated and performed by the GEOMAR Research Center for Marine Geosciences at Kiel and the College of Ocean and Atmospheric Sciences (COAS) of Oregon State University, Corvallis. We thank these institutions for the invitation to participate. The TECFLUX programme is supported in Germany by the Bundesministerium für Bildung und Forschung and in the United States by the National Science Foundation. We thank the officers, crew and shipboard scientific party of RV SONNE for support at sea during the expedition SO-143, P. Linke and M. Elvert for providing some of the sediment samples, A. Kähler, T. Lösekann, H. Löbner and G. Tsounis for help with the

SRR measurements, and J. Wulf for help with the micrograph photography. We thank C. Arnosti for corrections to the manuscript, and D. Valentine and W. Reeburgh for showing us their manuscript before publication.

Correspondence and requests for materials should be addressed to A.B. (e-mail: aboetius@mpi-bremen.de).

Modern freshwater microbialite analogues for ancient dendritic reef structures

Bernard Laval*[†], **Sherry L. Cady**[‡], **John C. Pollack**[§], **Christopher P. McKay**^{||}, **John S. Bird**^{*}, **John P. Grotzinger**[¶], **Derek C. Ford**[#] & **Harry R. Bohm**^{*}

* School of Engineering Science, Simon Fraser University, Burnaby, British Columbia, V5A 1S6, Canada

‡ Department of Geology, Portland State University, Portland, Oregon 97201, USA

§ Forest Sciences, Nelson Forest Region, British Columbia, V1L 4C6, Canada

|| NASA Ames Research Centre, Moffett Field, California 94035, USA

¶ Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

McMaster University, Hamilton, Ontario, L8S 4K1, Canada

Microbialites are organosedimentary structures that can be constructed by a variety of metabolically distinct taxa¹. Consequently, microbialite structures abound in the fossil record, although the exact nature of the biogeochemical processes that produced them is often unknown². One such class of ancient calcareous structures^{3–5}, *Epiphyton* and *Girvanella*, appear in great abundance during the Early Cambrian. Together with *Archeocyathids*, stromatolites and thrombolites, they formed major Cambrian reef belts. To a large extent, Middle to Late Cambrian reefs are similar to Precambrian reefs⁶, with the exception that the latter, including terminal Proterozoic reefs⁷, do not contain *Epiphyton* or *Girvanella*. Here we report the discovery in Pavilion Lake, British Columbia, Canada, of a distinctive assemblage of freshwater calcite microbialites, some of which display microstructures similar to the fabrics displayed by *Epiphyton* and *Girvanella*. The morphologies of the modern microbialites vary with depth, and dendritic microstructures of the deep water (>30 m) mounds indicate that they may be modern analogues for the ancient calcareous structures. These microbialites thus provide an opportunity to study the biogeochemical interactions that produce fabrics similar to those of some enigmatic Early Cambrian reef structures.

Pavilion Lake is nestled in a steep-walled limestone valley known as Marble Canyon⁸. It is a small (5.7 × 0.8 km, Fig. 1), deep (maximum recorded depth 65 m), slightly alkaline (pH 8) freshwater lake. The lake water is very clear with Secchi depths of more than 15 m. Surface streams do not enter the lake and the predominance of limestone canyon walls indicates that karst hydrology dominates. Multi-beam and side-scan sonar profiles indicate that the microbialite assemblages that are distributed along the walls of the lake basins and surrounded by white silt are orientated roughly perpendicular to the shoreline (Fig. 1, inset). A decrease with depth in both the friability and porosity of the Pavilion Lake microbialites accompanies the distinct changes in their morphologies and

[†] Present address: Centre for Water Research, University of Western Australia, Nedlands, Western Australia 6907, Australia

Table 1 Pavilion Lake uranium series analyses for dating microbialites

Sample	Height above base (cm)	Weight (gm)	U content (p.p.m.)	$^{234}\text{U}/^{238}\text{U}$ activity ratios	$^{230}\text{Th}/^{234}\text{U}$	$^{230}\text{Th}/^{232}\text{Th}$	Corrected age* (yr)	Uncorrected age† (yr)
PAV1	3.0	8.2	0.315	1.4339	0.1075	2	$4,430 \pm 1,433$	$12,300 \pm 1,400$
	7.5	8.5	0.316	1.4388	0.0781	3	$4,600 \pm 1,175$	$8,800 \pm 1,165$
	13.5	8.1	0.348	1.4387	0.0547	3	$2,570 \pm 1,140$	$6,100 \pm 1,000$
PAV 2	4.5	11.0	0.332	1.4281	0.0331	1	$2,240 \pm 1,000$	$3,650 \pm 860$
	11.0	8.2	0.222	1.4539	0.0787	2	$3,360 \pm 2,100$	$8,900 \pm 2,100$
PAV3	10 to 15	60	0.113	1.7331	0.0591	1.5	$2,300 \pm 2,000$	$6,600 \pm 2,700$

* Corrected ages calculated with the program TIMS Age 4U2U²².

† Uncorrected ages represent the maximum possible age.

internal microstructures. X-ray diffraction showed that all of the microbialite morphotypes (collected at depths of 15, 21, 24, 27 and 32 m) consist of calcite grains characterized by similar grain size and crystallinity.

The shallow to intermediate facies (about 10 m deep) consists of microbialite mounds that range from several centimeters to a few decimeters in height. These mounds comprise smaller clusters of discrete but interconnected round aggregates of calcite grains covered by photosynthetic microbial communities and their calcified remains (Fig. 2a). The porous microenvironment of these microbialite mounds traps additional debris that reaches the lake bottom and serves as a refuge for metazoan grazers (for example, nematodes, oligochaetes and midges). A thin (< 5 mm thick) green-pigmented, cyanobacteria-dominated microbial mat that contains coccoids (*Gloeocapsa* sp.; ~1–2 µm wide), rods (*Synechococcus* sp.; 0.5 µm wide) and tangentially orientated filaments (*Fischerella* sp. and *Pseudoanabaena* sp.; ~2 µm wide) covers the surfaces of these lithified but extremely friable mounds (Fig. 2c). Isolated communities of larger (~7 µm wide) purple-pigmented cyanobacteria filaments (*Oscillatoria* sp. and *Calothrix* sp.) occur on or between mat-rich areas as tiny tufts of erect filaments or as globular aggregates of intertwined filaments (Fig. 2b). Numerous heterotrophs and diatoms (*Gomphonema*, *Cyclotella* and *Achnanthes*) are also present. On the outer surfaces of the microbialites, we have found that the thinner cyanobacterial filaments are more susceptible to calcification than the larger filaments (Fig. 2d). Filamentous microbial fossils of both sizes, however, occur within the microbialites (Fig. 3a). Both populations contribute to microbialite accretion by trapping authigenic mineral precipitates that form in the water column. The calcification of picoplankton communities of cyanobacteria produces whitening events when populations

bloom^{9,10}, resulting in an accumulation of authigenic carbonate sediment around the microbialite fields, *Synechococcus* sp. ($\leq 82.5 \times 10^3$ cells ml⁻¹, J. Thompson, personal communication) being the dominant picoplankton in Pavilion Lake.

Large microbialite domes (up to 3 m high, Fig. 3b) found at intermediate depths (~20 m) also consist of discrete but interconnected carbonate clusters similar to those described above. These microbialites, however, consist of rather closely spaced aggregated clusters whose preferred orientation forms structural components with an appearance similar to that of vertically ribbed 'leaves'. Less friable than the shallow to intermediate depth microbialites, these separate but overlapping structures connect at their bases to adjacent leaves. Small cones occasionally found at the tops of these structures may reach a height of several centimeters, and their outer surfaces, like the surfaces of the microbialite leaves upon which they formed, consist of vertically orientated ridges of calcite globules (< 1 cm diameter). The microbialite structures shown in Fig. 3a resemble many modern lacustrine microbialites^{11–14}, as do those shown in Fig. 3b, although to a lesser extent.

At intermediate to deep depths (20–30 m) the microbialites display a morphology best described as a combination of cone-shaped and leaf-like structures (Fig. 3c). In this case, the individual leaves and cones are larger (20–35 cm in height), and noticeably more dense than those found at shallower depths. The cones of the intermediate- to deep-water microbialites often have one or more internal conduit up to 5 mm in diameter. Based on their appearance and the presence of internal conduits, it is probable that the distribution of the intermediate to deep, cone-topped microbialites corresponds to regions of groundwater seepage into the lake. These cone structures resemble the inactive tufa towers distributed along

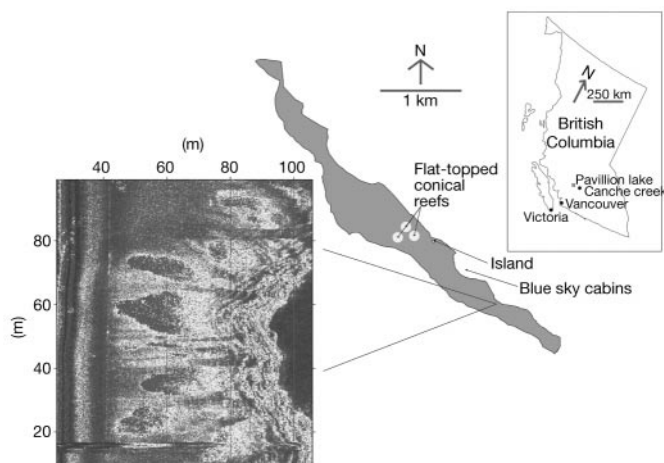


Figure 1 Unique assemblage of calcite microbialites, Pavilion Lake, British Columbia. The drawing of the lake shows the position of a ridge, topped by three flat, shallow reefs, that partitions the lake into two basins. Microbialites occur in the southern and northern basins, which have maximum depths of 45 and 65 m, respectively. Inset, side-scan sonar

image showing that microbialites (strong signals) occur along the edges of the basins in finger-like fields orientated roughly perpendicular to the shoreline. An unconsolidated calcareous sediment that produces relatively weak sonar returns surrounds the microbialites.

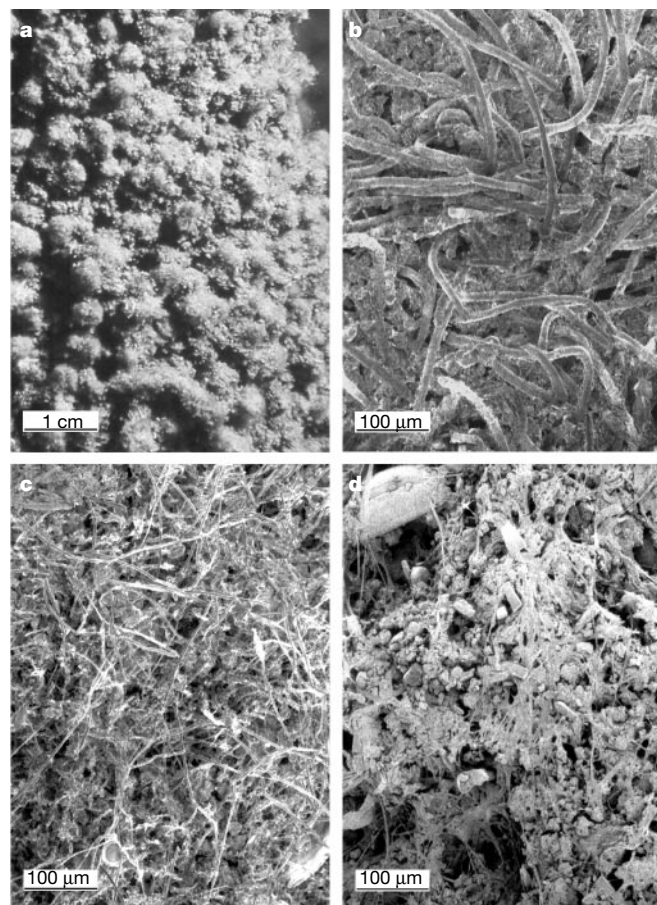


Figure 2 Photosynthetic microbial mat communities on microbialites. **a**, Surface of shallow-intermediate facies mound contains rounded tufts of larger, purple-pigmented filamentous cyanobacteria (*Oscillatoria* sp. and *Calothrix* sp.; shown in **b**) that are surrounded by green-pigmented mat of filamentous cyanobacteria (*Pseudoanabaena* sp. and *Fischerella* sp.; shown in **c**). **b**, Scanning electron microscopy (SEM) photomicrograph of uncalcified aggregates of purple-pigmented bacteria on microbialite surface. SEM specimens were fixed, dehydrated, critical point dried and Ir coated. **c**, Loose networks of filamentous green-pigmented cyanobacteria often bind trapped sediment. **d**, Calcification of the cyanobacteria filaments and their extracellular matrix at the surfaces of the microbialites perpetuates microbialite growth by providing a lithified substrate for the development of subsequent microbial mats.

the shores of Mono Lake¹⁵. Possible ancient cone-shaped vent structures with internal conduits include silicified stromatolites of probable lacustrine setting from the early Cambrian or Vendian age Antrim Plateau Volcanics¹⁶.

The deepest microbialite facies (> 30 m) consist of extremely dense structures that are sometimes topped by smaller cones that reach 5–10 cm in height (Fig. 3d). These microbialite structures display a unique morphology and internal dendritic microstructure that we propose are possible analogues for the ancient carbonate structures *Epiphyton* and *Girvanella*. Although the biological affinities of these ancient structures are uncertain owing to poor cellular preservation and the previous lack of a modern analogue, they are typically regarded as consisting of calcified cyanobacteria^{2,17}. We also note a similarity in gross morphology between some of the Pavilion Lake microbialites and the class of Proterozoic structures known as *Jacutophyton*¹⁸. However, the laminated internal fabric of the latter is not comparable to the dendritic bushy fabric characteristic of the modern, deep-water Pavilion Lake structures.

Geochemical analyses of water samples collected in October 1998, at depths of 12, 18, 24 and 30 m, indicate that in the autumn, before

any seasonal overturn, lake water below the thermocline is slightly undersaturated with respect to calcite. Direct measurements of photosynthetically active radiation (PAR) using a LiCor submersible quantum sensor (400–700 nm) has shown that at this time of year the amount of light drops by about one order of magnitude for each 10-m increase in depth. Given that the microbial mat populations associated with the microbialite structures in Pavilion Lake are dominated by oxygenic phototrophs, attenuation of PAR with increasing depth probably results in a decrease in the metabolic activity of the cyanobacteria in the autumn. We also observed differences as a function of depth in the taxonomic abundances and type of cyanobacteria in the microbial mats associated with the microbialites. On-going investigations are focused on determining how variations in the consortium populations and annual fluctuations in lake water geochemistry influence the development and preservation of microbial biosignatures (for example, microbialite morphologies and microstructures, microfossils and chemofossils) in these structures.

The microbialite structures grow out of precipitated carbonate silts that bury clastic sediments of postglacial age, and thus must be younger than the deglaciation (~11,500 yr ago) and the paraglacial phase of rapid clastic sedimentation that succeeded it. To establish approximate ages and rates of growth more precisely, six ²³⁰Th/²³⁴U analyses were made of the calcite in the lower to upper-middle parts of two whole mounds (collected from depths of 27 and 32 m) using thermal ionization mass spectrometry (TIMS)¹⁹, plus one bulk analysis by alpha spectrometry of the centre of a third (collected from 32 m depth). The mounds were 20–30 cm in height. Uranium content was 0.11–0.35 p.p.m. (Table 1). Owing to the organic framework trapping clay particles, the detrital thorium content was high in all cases, with ²³⁰Th/²³²Th ratios between 1 and 3; the true bases (buried within the silts) were too dirty to attempt analysis. The detrital thorium introduces considerable potential error into the calculation of the ages. However, the U-series data are internally consistent, and similar for all three specimens. Raw ages that range from 12,300 ± 1,400 to 3,650 ± 860 yr represent the maximum possible age. Corrected U–Th ages, due to detrital contamination, will be younger than the uncorrected ages. Mound PAV1 displays a linear relationship between height above its base and ²³⁰Th/²³⁴U (three TIMS analyses). Extrapolation indicates that it began to grow about 6,000 years ago and has extended at rates between 2.5 and 3.0 cm per thousand years. The other two mounds probably have similar histories.

The deep-water structures, a possible modern analogue of *Epiphyton* and *Girvanella*, display a dendritic microstructure composed of bushes comprising upwardly splaying aggregates of calcified microorganisms (Fig. 4a). Large calcified filaments (~10 µm diameter), the principal architectural component of these dendritic bushes, along with thinner calcified filaments (~2 µm diameter), which occur alongside the larger microfossils, typify the bimodal size distribution observed in the viable filament populations that dominate the microbial mat communities. Dendritic branches, in particular, exhibit textural evidence that mimics the arrangement of viable purple-pigmented filaments in the tufts (Fig. 4b). Although *Oscillatoria* sp. and *Calothrix* sp., located in the tufts on the outer microbialite surfaces, were not visibly calcified, the presence inside the dendrites of calcified microbial fossils that are only a few micrometres thicker than the viable filamentous organisms, indicates their calcification begins within the microbial mats, rather than at the outer mat surface. Microbial calcification induced by the metabolic activity of heterotrophic bacteria within the aphotic zone of mats has been proposed as a significant factor in the formation of mesoclotted fabrics of thrombolites and in having produced the micritic, clotted fabrics exhibited by many ancient microbialites²⁰. However, it is not known whether calcification of the ancient analogues occurred as a result of postmortem decomposition of sheaths by precipitation-inducing heterotrophic bacteria²¹, or if it

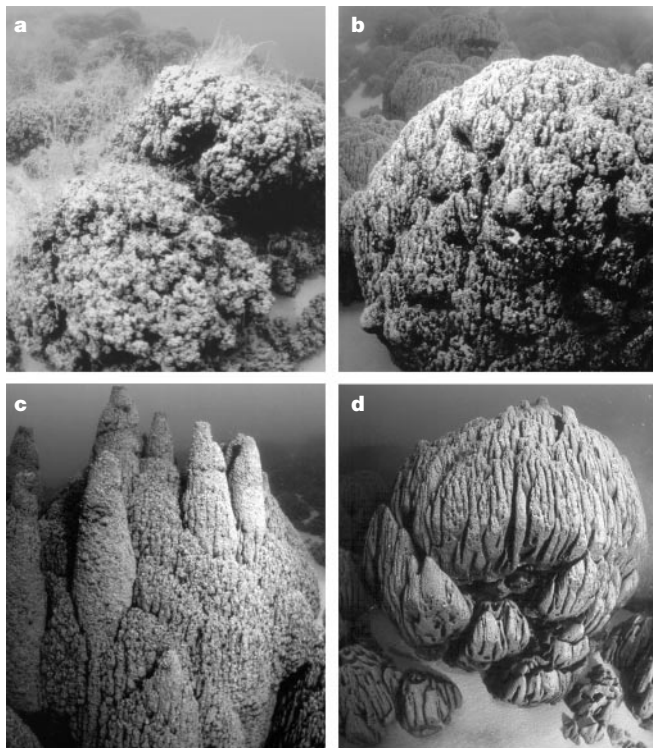


Figure 3 Pavilion Lake microbialite morphotypes. **a**, Shallow-intermediate facies mounds (10–15 m depth; centimetres to decimetres in diameter) serve as substrates for benthic freshwater lake flora (charophytes). **b**, Calcified microbial remains and trapped sediment comprise 'leaves' of intermediate facies domes (~20 m deep; decimetres to metres in diameter). **c**, Cone-shape seepage structures with hollow internal conduits that open at the tops of the cones on the intermediate depth microbialites (~20–30 m; decimetres to metres in diameter). **d**, Relatively dense calcite leaves of deep water mounds (> 30 m; centimetres to metres in diameter) display unique dendritic microstructure.

occurred as a vital effect². Our findings underscore the unique opportunities the Pavilion Lake ecosystem presents in elucidating microbialite morphogenesis and microbial calcification processes, and in improving our ability to interpret properly the fabrics of ancient carbonates. □

Received 25 April; accepted 10 August 2000.

1. Burne, R. V. & Moore, L. S. Microbialites: organosedimentary deposits of benthic microbial communities. *Palaio* **2**, 241–254 (1987).
2. Riding, R. in *Calcareous Algae and Stromatolites* (ed. Riding, R.) 55–87 (Springer, Berlin, 1991).
3. Ahr, W. M. Paleoenvironment, algal structures, and fossil algae in the Upper Cambrian of Central Texas. *J. Sedim. Petrol.* **41**, 205–219 (1971).
4. James, N. P. Megablocks of calcified algae in the Cow Head Breccia, western Newfoundland: Vestiges of a Cambro-Ordovician platform margin. *GSA Bull.* **92**, 799–811 (1981).
5. Read, J. F. & Pfeil, R. W. Fabrics of allochthonous reefal blocks, Shady dolomite (Lower to Middle Cambrian), Virginia Appalachians. *J. Sedim. Petrol.* **53**, 761–778 (1983).
6. Grotzinger, J. P. in *Controls on Carbonate Platform and Basin Development* Vol. 44 (eds Crevello, P. D., Wilson, J. L., Sarg, J. F. & Read, J. F.) 79–106 (SEPM Special Publication, Tulsa, 1989).
7. Grotzinger, J. P. & Khetani, A. in *Abstracts with Programs (Northeast Section) Conference Proceedings* Vol. 27 (GSA Publications, Denver, 1995).
8. Fischl, P. *Limestone and Dolomite Resource in B. C.* Open File 1992-18 (British Columbia Ministry of Energy, Mines and Petroleum Resources, Geological Survey Branch, Victoria, 1992).
9. Thompson, J. B. & Ferris, G. F. Cyanobacterial precipitation of gypsum, calcite, and magnesite from natural alkaline lake water. *Geology* **18**, 995–998 (1990).
10. Thompson, J. B., Schultze-Lam, S., Beveridge, T. J. & Des Marais, D. J. Whiting events: Biogenic origin due to the photosynthetic activity of cyanobacteria picoplankton. *Limnol. Oceanogr.* **42**, 133–141 (1997).
11. Halley, R. B. in *Stromatolites* (ed. Walter, M. R.) 435–445 (Elsevier, Amsterdam, 1976).
12. Eggleston, J. R. & Dean, W. E. in *Stromatolites* (ed. Walter, M. R.) 479–488 (Elsevier, Amsterdam, 1976).
13. Thompson, J. B., Ferris, F. G. & Smith, D. A. Geomicrobiology and sedimentology of the mixolimnion and chemocline in Fayetteville Green Lake, New York. *Palaio* **5**, 52–75 (1990).
14. Ferris, F. G., Thompson, J. B. & Beveridge, T. J. Modern freshwater microbialites from Kelly Lake,

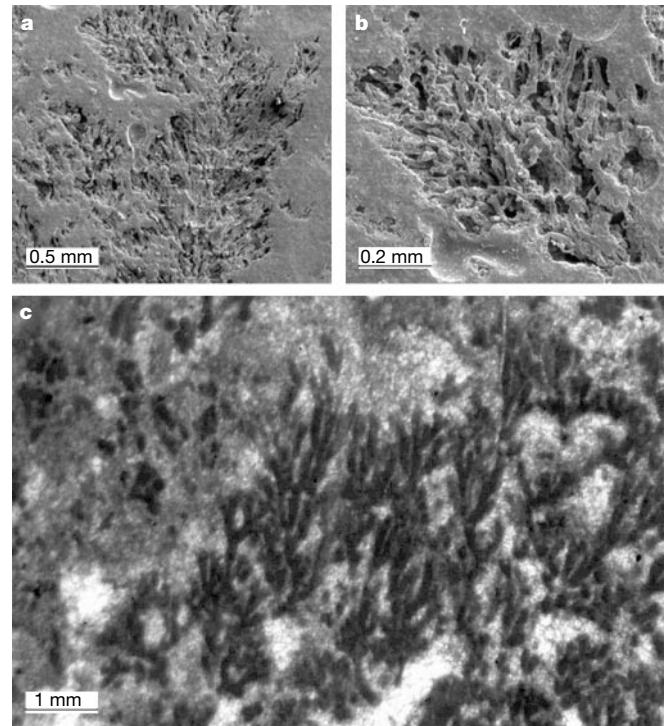


Figure 4 Modern microbialite microstructure compared to fabric of ancient carbonate deposit. **a**, **b**, Modern microbialite microstructure; **c**, ancient carbonate deposit. **a**, SEM shows individual bushes of calcified filamentous microfossils and trapped sediment. Authigenically calcified picoplankton comprise calcareous sediment matrix. SEM specimens were impregnated in epoxy resin, etched in 2M HCl, and Ir coated. **b**, Branches of dendrite consist of radiating splays of larger calcified filaments, supported by a framework of thinner calcified filaments, calcified extracellular matrix, and trapped and bound sediment particles. **c**, Optical thin section shows dendritic fabric of potential ancient analogue *Epiphyton*. In the early Cambrian Shady dolomite of southwest Virginia⁵, *Epiphyton* forms mound-like structures similar to the large mounds found in Pavilion Lake.

- British Columbia, Canada. *Palaio* **12**, 213–219 (1997).
15. Bischoff, J. L., Stine, S., Rosenbauer, R. J., Fitzpatrick, J. A. & Stafford, T. W. Ikaite precipitation by mixing of shoreline springs and lake water, Mono Lake, California, USA. *Geochim. Cosmochim. Acta* **57**, 3855–3865 (1993).
 16. Walter, M. R. Stromatolites and the biostratigraphy of the Australian Precambrian and Cambrian. *Spec. Pap. Palaeontol.* **11**, 103–104 (1972).
 17. Wray, J. L. *Calcareous Algae* (Elsevier, Amsterdam, 1977).
 18. Bertrand-Sarfati, J. & Moussine-Pouchkine, A. Evolution and environmental conditions of *Conophyton-Jacotophyton* associations in the Atar dolomite (upper Proterozoic, Mauritania). *Precamb. Res.* **29**, 207–234 (1985).
 19. Edwards, R. L., Chen, J. H. & Wasserburg, G. J. ²³⁸U–²³⁴U–²³²Th–²³²Th systematics and the precise measurement of time over the past 500,000 years. *Earth Planet. Sci. Lett.* **81**, 175–192 (1986).
 20. Chafetz, H. S. & Buczynski, C. Bacterially induced lithification of microbial mats. *Palaio* **7**, 277–293 (1992).
 21. Pratt, B. *Epiphyton and Renalcis*—diagenetic microfossils from calcification of coccoid blue-green algae. *J. Sedim. Petrol.* **54**, 948–971 (1984).
 22. Lauritzen, S. -E. & Lundberg, J. Calibration of the speleothem delta function: an absolute temperature record for the Holocene in northern Norway. *Holocene* **9**, 659–669 (1999).

Acknowledgements

We thank the Ts'kw'aylaxw First Nation Tribe and Ocean Pacific Water Sports for their logistical assistance and support. Underwater photographs are courtesy of D. Reid, Ocean Photography. D. F. Blake, L. Duncan and B. Schauer provided technical support. M. Taylor, C. Stack and C. R. Omelon assisted with the identification of cyanobacteria and diatoms. Protection for the unique microbialites in Pavilion Lake is underway through the Protected Areas Strategy (PAS) Goal 2 initiative of the British Columbia provincial government.

Correspondence and requests for materials should be addressed to S.L.C. (e-mail: cadys@pdx.edu).

Minimization of Boolean complexity in human concept learning

Jacob Feldman

Department of Psychology, Center for Cognitive Science, Rutgers University, New Brunswick, New Jersey, 08903, USA

One of the unsolved problems in the field of human concept learning concerns the factors that determine the subjective difficulty of concepts: why are some concepts psychologically simple and easy to learn, while others seem difficult, complex or incoherent? This question was much studied in the 1960s¹ but was never answered, and more recent characterizations of concepts as prototypes rather than logical rules^{2,3} leave it unsolved^{4–6}. Here I investigate this question in the domain of Boolean concepts (categories defined by logical rules). A series of experiments measured the subjective difficulty of a wide range of logical varieties of concepts (41 mathematically distinct types in six families—a far wider range than has been tested previously). The data reveal a surprisingly simple empirical ‘law’: the subjective difficulty of a concept is directly proportional to its Boolean complexity (the length of the shortest logically equivalent propositional formula)—that is, to its logical incompressibility.

How human learners extract rules from patterns of data, and the relative subjective complexity of different kinds of rules, is central to investigations of human learning. When classifying novel stimuli, humans tend to seek simple rules rather than prototypes⁷. Subjective complexity has a central theoretical role in popular rule-plus-exception models of concept learning⁸, and more generally in debates on the nature of learning^{9,10}, where it is often argued that examples failing to obey some ‘simple’ primary rule might be stored in an alternative manner, for example, verbatim. Earlier research focused on an extremely limited variety of logical forms, virtually exclusively comprising concepts defined with only two features. Most research considered just two types of logical rule, conjunction and disjunction, with one empirical result—that conjunctive (‘and’) concepts are learned more easily than disjunctive (‘or’) ones—dominating discussion. One early report¹¹ suggested a role for logical complexity, but this proposal was never followed up. Most theoretical accounts since then^{1,8,12,13} have taken the subjective preference for conjunctive concepts as axiomatic; usually this is the only principle invoked pertaining to logical form. But no independently motivated explanation of it has ever emerged.

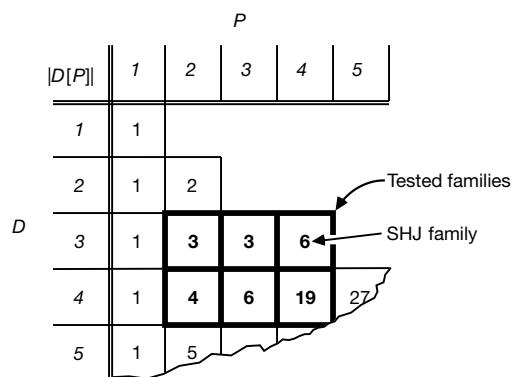


Figure 1 The $D[P]$ hierarchy, showing the number of cases ($D[P]$) in each family. The hierarchy extends infinitely to all values of D , and all values of P less than or equal to 2^{D-1} . Families in the bold box are those tested in the experiments. Family 3[4] is the one originally considered by Shepard *et al.*¹⁴.

One study, that of Shepard, Hovland and Jenkins¹⁴ investigated concepts with three features. Their stimulus classification and basic empirical result provides the motivation for my study. Shepard *et al.* considered Boolean concepts defined over three features with four positive and four negative examples. Such concepts fall into six logical types, designated I–VI (the SHJ types). All concepts with three features and four positive examples are isomorphic to one of the six types, but no example of one type is isomorphic to one of a different type, making this a complete classification. Illustrations of the six types can be found in Table 1 in the subtable labelled 3[4] (the notation is explained below). Shepard *et al.* found that the types differed in subjective difficulty in the order $I < II < [III, IV, V] < VI$, with types III, IV and V having approximately equal difficulty. This basic empirical pattern has been much discussed and occasionally replicated since^{8,15}, but while it has been modelled by simulations involving many free parameters, it has resisted any simple or elegant theoretical explanation.

When the SHJ types are considered from the perspective of mathematical logic, however, a simple explanation of the difficulty ordering emerges: the difficulty of the six types is precisely predicted by their Boolean complexity. The Boolean complexity of a propositional concept is the length of the shortest Boolean formula logically equivalent to the concept, usually expressed in terms of the number of literals (positive or negative variables)^{16,17}. (For convenience, I write $a \wedge b$ as ab , $a \vee b$ as $a + b$, and $\neg a$ as a' .) For example, the concept $ab + ab'$ is equivalent to $a(b + b')$ and thus to a , and hence has Boolean complexity 1; whereas $ab + a'b'$ has no shorter equivalent, and hence has Boolean complexity 4. Like Kolmogorov complexity^{18–21}, Boolean complexity is an essentially universal measure of the intrinsic mathematical complexity or ‘incompressibility’ of the propositional concept¹⁶. Finding the shortest formula

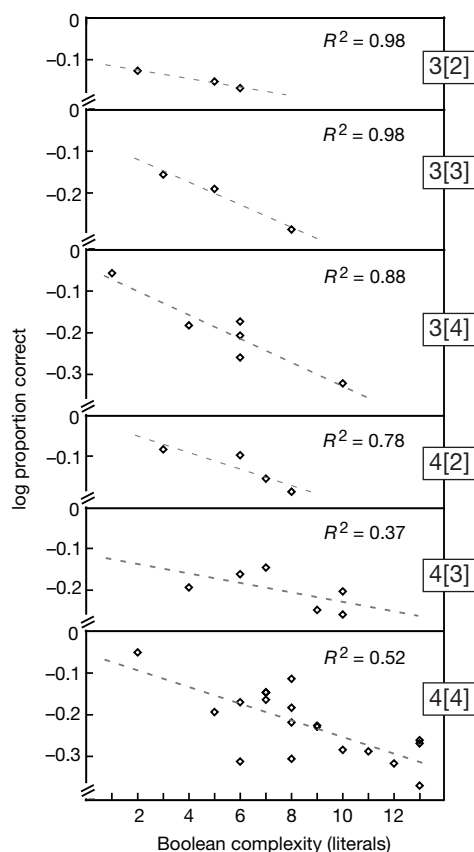


Figure 2 Results, showing log proportion correct as a function of Boolean complexity for each family (collapsing over parity). Full plots including all cases in both parities can be found in Supplementary Information.

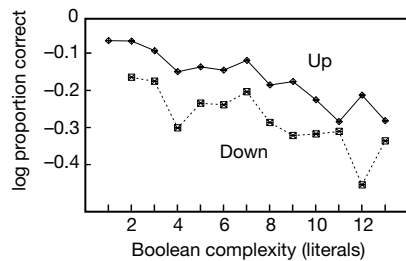


Figure 3 Results by parity, collapsing over family. The 3[4] family is included with the Up cases. These results should be interpreted with caution because the different families may have different overall levels of difficulty (controlled in part by the amount of learning time, and only approximately equalized by the chosen levels).

equivalent to a given formula is computationally intractable²², and in practice Boolean reduction can only be achieved through approximate computational techniques (such as factorization, as illustrated in the example).

Now consider an arbitrary Boolean concept defined by P positive examples over D binary features (for example, in the six SHJ types, $D = 3$ and $P = 4$). Any such concept is logically equivalent to the disjunction of its P constituent objects, each of which is a conjunction of D features, a form known as a disjunctive normal formula (DNF). Hence, each such concept is equivalent to a DNF with DP literals. For example, each of Shepard *et al.*'s six types is equivalent to a formula of length 12 ($= 3 \times 4$). The DNF is the completely 'uncompressed' form; it in effect lists verbatim all the objects that satisfy the concept. When each of the SHJ types is compressed as much as possible (again, using heuristic techniques), they have lengths 1, 4, 6, 6, 6, and 10, respectively (Table 1, section 3[4], gives

all the full DNFs and minimal formulae). These Boolean complexity values predict the order of empirical difficulty precisely. This exact correspondence has not previously been noted, though Shepard *et al.* speculated about it in their original paper¹⁴, and the relation between Boolean complexity and human learning has never been comprehensively tested.

To investigate more thoroughly the role of Boolean complexity in learning, the next natural step is to extend the range of propositional concepts beyond Shepard *et al.*'s original six. Those six cases exhaust all possibilities with $D = 3$ and $P = 4$; an obvious extension is to consider different values of D , P , or both. Such changes lead to completely different classifications. For example, with $D = 3$ and $P = 3$ there are three cases, each necessarily distinct from any of the six used by Shepard *et al.*

Varying both D and P defines a parametric system of families, each analogous to, but different from, the SHJ family. I use the notation $D[P]$ to indicate the family of concept types involving D features and P positive examples, and denote the cases within that family by $I_{D[P]}$, $II_{D[P]}$ and so on. Under this notation, the SHJ family is 3[4], and its six cases are $I_{3[4]}$, $II_{3[4]}$, ..., $VI_{3[4]}$. Figure 1 gives the number of cases $|D[P]|$ for each family from $D = 1$ to 4 and for $P = 1$ to 2^{D-1} . Values of P larger than 2^{D-1} duplicate smaller values; for example, a concept with three positives and five negatives is simply the converse of one with five positives and three negatives, and hence, has essentially the same form and the same complexity.

The $D[P]$ hierarchy is completely comprehensive, in that it encompasses the entire universe of propositional formulae. All Boolean concepts fall into a family that appears somewhere in (the infinite extension of) Fig. 1, and all Boolean concepts are logically equivalent to one somewhere on the (infinite extension of) Table 1. Hence, the $D[P]$ complexity hierarchy is the natural testbed in which to evaluate any psychological theory that makes reference

Table 1 Details of two concept families tested in the experiments

Family*	Case	DNF	Minimal formula	Complexity	Illustration
3[3]	$I_{3[3]}$	$a'b'c' + a'b'c + a'bc'$	$a'(bc)'$	3	
	$II_{3[3]}$	$a'b'c' + a'b'c + abc'$	$a'b' + abc'$	5	
	$III_{3[3]}$	$a'b'c' + a'bc + ab'c$	$a'(b'c' + bc) + ab'c$	8	
3[4]	$I_{3[4]}$	$a'b'c' + a'b'c + a'bc' + a'bc$	a'	1	
	$II_{3[4]}$	$a'b'c' + a'b'c + abc' + abc$	$ab + a'b'$	4	
	$III_{3[4]}$	$a'b'c' + a'b'c + a'bc' + ab'c$	$a'(bc)' + ab'c$	6	
	$IV_{3[4]}$	$a'b'c' + a'b'c + a'bc' + ab'c'$	$a'(bc)' + ab'c'$	6	
	$V_{3[4]}$	$a'b'c' + a'b'c + a'bc' + abc$	$a'(bc)' + abc$	6	
	$VI_{3[4]}$	$a'b'c' + a'bc + ab'c + abc'$	$a(b'c + bc') + a'(b'c' + bc)$	10	

For each type the Roman-numeral case label, disjunctive normal formula, minimal formula, Boolean complexity, and a schematic illustration of the concept as a set of vertices in Boolean D -space are shown. Minimal formulae and complexity values are derived using heuristic minimization techniques, such as factoring (for details see Supplementary Information). All concepts with D features and P positives can be created by taking one of these cases and permuting and/or inverting the axes in Boolean D -space—equivalent to permuting the features and the parity of each feature. Ordering of cases within each family is arbitrary.

* Family 3[4] is the original family from Shepard *et al.*¹⁴. For details of other families see Supplementary Information.

to the logical structure of concepts.

The experiments below focus on six families: 3[2], 3[3], 3[4] (the SHJ family), 4[2], 4[3] and 4[4]; that is, all concepts with three or four features and between two and four positive examples, comprising 41 cases in the six families. Table 1 gives details of families 3[3] and 3[4]. Each concept (except in 3[4], where positive and negative sets are the same size) was tested in two versions: an Up version, in which the smaller set of objects was labelled 'positive', and a Down version, in which the larger set was labelled 'positive'. This factor, which is orthogonal to Boolean complexity (because every concept ϕ and its complement ϕ' have the same complexity) will be referred to as parity.

Results are plotted in Fig. 2 (separated by family) and Fig. 3 (separated by parity). The main trend is that performance decreases monotonically with increasing Boolean complexity, with an approximately constant advantage for Up versus Down cases. Together the complexity and parity effects explain about half of the variance ($R^2 = 0.5017$, $F(2, 73) = 36.76$, $P < 0.0001$). The parity effect, which is clearly visible in all families, has antecedents in the literature^{12,23–25}. Again, this factor is orthogonal to complexity and thus cannot be explained by complexity minimization.

My main conclusion is that subjective conceptual difficulty is well predicted by Boolean complexity. For each concept, learning is successful to the degree that the concept can be faithfully compressed^{26,27}. However, complexity and parity alone do not provide a complete account; there is substantial residual variance in the data (especially in higher-dimensional cases, and in Down cases in all dimensions) suggesting the influence of some unknown processing factors and strategies, which require future study.

The significance of these results is best viewed from an historical perspective. Concept learning research in the 1960s was largely preoccupied by the dichotomy between conjunctive and disjunctive concepts. In retrospect, much of the interest in this comparison derived from the fact that it seemed to imply a divergence between logical complexity, in which the two types are manifestly equal, and psychological complexity, in which the two types differ markedly. The contrast suggested a predominant role of extra-logical, and hence perhaps more subjective and mysterious, factors.

But in the light of the current study, this conclusion seems to have been premature. In the $D[P]$ classification, conjunction ab and disjunction $a + b$ represent the same case, $I_{2[1]}$; conjunction is the Up version (one positive) and disjunction is the Down version (three positives). That is, conjunction and disjunction are isomorphic, in that each divides Boolean 2-space into one vertex versus the other three; in conjunction the one is labelled positive and the rest negative, in disjunction the reverse. (Conditional $a \rightarrow b$ is also of the same type.) All other propositional concepts studied in the literature (again, apart from ref. 14) are bivariate concepts falling into the 2[2] family, which has only two cases, each appearing in two complementary versions: affirmation a and negation a' (both $I_{2[2]}$); and biconditional $a \leftrightarrow b$ (that is, $a'b' + ab$) and exclusive disjunction $ab' + a'b$ (both $II_{2[2]}$). Conjunction/disjunction, affirmation/negation and biconditional/xor have Boolean complexities 2, 1 and 4, respectively, approximately agreeing with their empirical difficulty ordering¹ once parity is taken into account. (This remark simply recapitulates Neisser and Weene's paper¹¹ in modern terminology; perhaps their paper would have had more influence if parity had been recognized as an orthogonal factor to complexity, thus improving agreement with the data.) In this light the subjective advantage of conjunction over disjunction admits a completely different explanation: the preference for Up over Down parity. If this account is correct, it places a very different interpretation on the main research result of this literature. In effect the traditional argument that logical complexity does not predict subjective difficulty was based on a few relatively simple cases that sit at the tip of a very large iceberg. The rest of the iceberg—the full $D[P]$ hierarchy—tells a very different story.

In a sense, this final conclusion may seem negative: human conceptual difficulty reflects intrinsic mathematical complexity after all, rather than some idiosyncratic and uniquely human bias. The positive corollary though is certainly more fundamental: subjective conceptual complexity can be numerically predicted and perhaps explained. The next step is to reconcile this finding with contemporary theory on concepts involving prototypes or other more realistic types of category. □

Methods

Concepts were presented to subjects using a fixed set of Boolean features in a world of 'amoebas' defined by simple binary features (shape of the nuclei, size of the nuclei, shading of the nuclei and number of nuclei). Subjects were instructed that in each block they were to view examples of a new species of amoeba, and that they were to try to learn to distinguish examples from non-examples. For each concept, the computer first generated a random concept of the desired type (by beginning with the normal form and then randomly permuting the assignment of features and the sign of each feature; hence all features played all roles in the logical forms roughly equally often). The computer then showed a screen displaying all positive and negative examples, which the subject was allowed to view for a fixed duration (always set at 5 Ps, for example 10 s in the 3[2] case, to make the six families of roughly equal overall difficulty). In the training screen, the positive examples appeared in the upper half of the screen labelled 'Examples', and all the rest of the 2^D objects appeared in the lower half of the screen labelled 'NOTexamples'. In the Up cases, there were P positive examples and $2^D - P$ non-examples, whereas in the Down cases there were $2^D - P$ positive examples and P negative examples. After the training screen, the subject was presented with a sequence of all 2^D objects in random order and asked to indicate with a button press whether the object was an example or not an example of the learned species. The computer recorded the proportion correct for each concept. At the end of the series of 2^D test trials, the computer would proceed to the training screen for the next concept.

There were 20, 22, 27, 44, 45 and 57 subjects in the six families, respectively. Each subject viewed concepts from only one $D[P]$ family, viewing all $|D[P]|$ concepts in both parities (except for 3[4], which has only one parity), with the entire set being repeated with new random concepts 8, 8, 3, 3, 2 and 1 times, respectively (these numbers chosen to make the total lengths of the sessions approximately equal and maximize the number of datapoints per subject).

Received 22 May; accepted 5 July 2000.

- Bourne, L. E. Knowing and using concepts. *Psychol. Rev.* **77**, 546–556 (1970).
- Posner, M. I. & Keele, S. W. On the genesis of abstract ideas. *J. Exp. Psychol.* **77**, 353–363 (1968).
- Rosch, E. H. Natural categories. *Cogn. Psychol.* **4**, 328–350 (1973).
- Osherson, D. N. & Smith, E. E. On the adequacy of prototype theory as a theory of concepts. *Cognition* **9**, 35–58 (1981).
- Medin, D. L. & Smith, E. E. Concepts and concept formation. *Annu. Rev. Psychol.* **35**, 113–138 (1984).
- Komatsu, L. K. Recent views of conceptual structure. *Psychol. Bull.* **112**, 500–526 (1992).
- Medin, D. L., Wattenmaker, W. D. & Hampson, S. E. Family resemblance, conceptual cohesiveness, and category construction. *Cogn. Psychol.* **19**, 242–279 (1987).
- Nosofsky, R. M., Palmeri, T. J. & McKinley, S. C. Rule-plus-exception model of classification learning. *Psychol. Rev.* **101**, 53–79 (1994).
- Pinker, S. Rules of language. *Science* **253**, 530–535 (1991).
- Smith, E. E., Patalano, A. L. & Jonides, J. Alternative strategies of categorization. *Cognition* **65**, 167–196 (1998).
- Neisser, U. & Weene, P. Hierarchies in concept attainment. *J. Exp. Psychol.* **64**, 640–645 (1962).
- Bourne, L. E., Ekstrand, B. R. & Montgomery, B. Concept learning as a function of the conceptual rule and the availability of positive and negative instances. *J. Exp. Psychol.* **82**, 538–544 (1969).
- Medin, D. L., Wattenmaker, W. D. & Michalski, R. S. Constraints and preferences in inductive learning: an experimental study of human and machine performance. *Cogn. Sci.* **11**, 299–339 (1987).
- Shepard, R., Hovland, C. L. & Jenkins, H. M. Learning and memorization of classifications. *Psychol. Monogr. Gen. Appl.* **75**, 1–42 (1961).
- Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C. & Glauthier, P. Comparing models of rule-based classification learning: a replication and extension of Shepard, Hovland and Jenkins (1961). *Mem. Cogn.* **22**, 352–369 (1994).
- Wegener, I. *The Complexity of Boolean Functions* (Wiley, Chichester, 1987).
- Givone, D. D. *Introduction to Switching Circuit Theory* (McGraw Hill, New York, 1970).
- Chaitin, G. J. On the length of programs for computing finite binary sequences. *J. Assoc. Comput. Machin.* **13**, 547–569 (1966).
- Kolmogorov, A. N. Three approaches to the quantitative definition of information. *Problems Inf. Transm.* **1**, 1–7 (1965).
- Solomonoff, R. J. A formal theory of inductive inference. Part I. *Inf. Control.* **7**, 1–22 (1964).
- Li, M. & Vitányi, P. *An Introduction to Kolmogorov Complexity and its Applications* (Springer, New York, 1997).
- Garey, M. R. & Johnson, D. S. *Computers and Intractability: A Guide to the Theory of NP-completeness* (Freeman, New York, 1979).
- Hovland, C. I. & Weiss, W. Transmission of information concerning concepts through positive and negative instances. *J. Exp. Psychol.* **45**, 175–182 (1953).
- Haygood, R. C. & Devine, J. V. Effects of composition of the positive category on concept learning. *J. Exp. Psychol.* **74**, 230–235 (1967).
- Nahinsky, I. D. & Slaymaker, F. L. Use of negative instances in conjunctive concept identification. *J. Exp. Psychol.* **84**, 64–68 (1970).
- Rissanen, J. Modeling by shortest data description. *Automatica* **14**, 465–471 (1978).

27. Chater, N. Reconciling simplicity and likelihood principles in perceptual organization. *Psychol. Rev.* **103**, 566–581 (1996).

Supplementary information is available on Nature's World-Wide Web site and as paper copy from the London editorial office of Nature.

Acknowledgements

Supported in part by a grant from the National Science Foundation. I thank S. J. Hanson, W. A. Richards, J. Tenenbaum and P. Tremoulet for helpful comments, and M. Balaban, E. Barenholtz, D. Berse, N. Folsom-Kovarik, J. Sutton and P. Tremoulet for assistance in data collection.

Correspondence and requests for materials should be addressed to J. F. (e-mail: jacob@ruccs.rutgers.edu).

Skin abnormalities generated by temporally controlled RXR α mutations in mouse epidermis

Mei Li[†], Arup Kumar Indra[†], Xavier Warot^{*}, Jacques Brocard^{*}, Nadia Messaddeq^{*}, Shigeaki Kato[‡], Daniel Metzger^{*} & Pierre Chambon^{*}

^{*} Institut de Génétique et de Biologie Moléculaire et Cellulaire, CNRS/INSERM/ULP, Collège de France, BP 163, 67404 Illkirch Cedex, France

[‡] Institute of Molecular & Cellular Biosciences, University of Tokyo, Yayoi, Bunkyo-ku, Tokyo 113, Japan

[§] CREST, Japan Science and Technology, Kawaguchi, Saitama 332, Japan

[†] These authors contributed equally to this work

Nuclear receptors for retinoids (RARs) and vitamin D (VDR), and for some other ligands (TRs, PPARs and LXRs), may be critical in the development and homeostasis of mammalian epidermis^{1–8}. It is believed that these receptors form heterodimers with retinoid X receptors (RXRs) to act as transcriptional regulators^{9,10}. However, most genetic approaches aimed at establishing their physiological functions in the skin have been inconclusive owing either to pleiotropic effects and redundancies between receptor isotypes in gene knockouts, or to equivocal interpretation of dominant-negative mutant studies in transgenic mice^{11,13–15}. Moreover, knockout of RXR α , the main skin RXR isotype, is lethal *in utero* before skin formation^{11,12,16,17}. Here we have resolved these problems by developing an efficient technique to create spatio-temporally controlled somatic mutations in the mouse. We used tamoxifen-inducible Cre–ER^T recombinases^{18,19} to ablate RXR α selectively in adult mouse keratinocytes. We show that RXR α has key roles in hair cycling, probably through RXR/VDR heterodimers, and in epidermal keratinocyte proliferation and differentiation.

To ablate RXR α in epidermis, we engineered mice carrying LoxP-site-containing (floxed) RXR α ^{L2} alleles (Fig. 1a) and used the K5–Cre–ER^T transgenic line in which tamoxifen (Tam) efficiently induces Cre-mediated recombination in basal layer keratinocytes¹⁹. K5–Cre–ER^{T(tg/tg)}/RXR α ^{L2/L2} mice mated with RXR α ^{+/-} (Fig. 1a; ref. 16) or RXR α ^{L2/+} mice yielded 'pro-mutant' mice hemizygous (tg/0) for K5–Cre–ER^T and carrying either one RXR α ^{L2} and one RXR α null (–) allele (K5–Cre–ER^{T(tg/0)}/RXR α ^{L2/-} genotype) or two L2 alleles (K5–Cre–ER^{T(tg/0)}/RXR α ^{L2/L2} genotype). At 14 weeks old, the pro-mutant mice were treated with Tam (5 days, 1 mg per day), and then retreated 2, 4 and 6 weeks later. Six weeks after the first Tam treatment (AFT), 80% of RXR α ^{L2} alleles were converted into RXR α ^{L-} alleles in the epidermis of mice carrying one or two floxed alleles (Fig. 1b). By 12 weeks AFT, almost all RXR α ^{L2} alleles had been

converted (Fig. 1b). As expected¹⁹, no RXR α disruption occurred in vehicle (oil)-treated mice (data not shown) and Cre-mediated excision of RXR α exon 4 was restricted to epidermis and some epithelia in which the K5 promoter is also active (for example, tongue, salivary gland, oesophagus; Fig. 1c).

Interestingly, hair loss (alopecia) was observed 6–7 weeks AFT in the ventral region of pro-mutant mice, but not in oil-treated pro-mutant mice or in Tam-treated K5–Cre–ER^{T(tg/0)}/RXR α ^{L2/+} 'control' littermates (data not shown). At 12–16 weeks AFT, large regions of ventral skin and smaller regions of dorsal skin were hairless (Fig. 2a, b; and data not shown). Cysts became visible under the skin surface and these enlarged and spread all over the body with time (Fig. 2c; and data not shown). With increasing age (> 20 weeks AFT), minor focal lesions appeared on hairless dorsal skin, on chins and behind ears (Fig. 2d; and data not shown). These were not caused by fights and were formed of crusts on top of hyperproliferative epidermis and inflammatory dermis (see below).

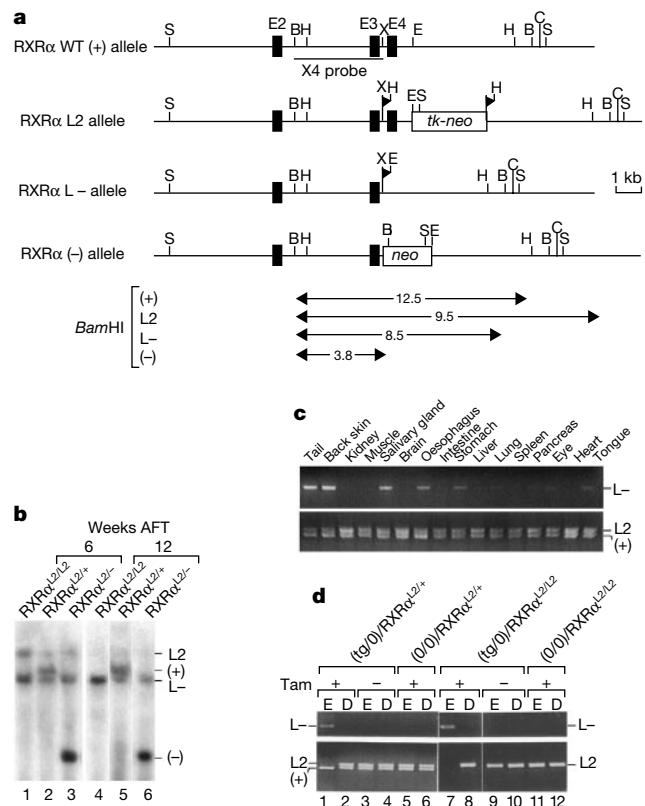


Figure 1 Tamoxifen-induced RXR α null mutation in adult mouse epidermis mediated by Cre–ER^T. **a**, Diagram of the wild-type RXR α genomic locus (+), the floxed RXR α L2 allele, the RXR α L- allele obtained after Cre-mediated excision of exon 4 (encoding the DNA-binding domain), and the RXR α null allele (–)¹⁶. Black boxes indicate exons (E2–E4). Restriction enzyme sites and probe X4 location are indicated. BamHI fragments are in kilobases (kb). B, BamHI; C, ClaI; E, EcoRI; H, HindIII; S, SpeI; X, XbaI. Arrowheads in L2 and L- alleles indicate LoxP sites. **b**, Tamoxifen (Tam)-induced generation of K5–Cre–ER^T-mediated RXR α ^{L-} alleles illustrated by Southern blot analysis of epidermal DNA isolated 6 (lanes 1–3) and 12 (lanes 4–6) weeks after the first Tam (1 mg) injection series (AFT). All mice were K5–Cre–ER^{T(tg/0)} and the RXR α genotypes are indicated. BamHI-digested DNA fragments corresponding to RXR α (+), L2, L- and (-) alleles are displayed. **c**, Tissue-specificity of Cre–ER^T-mediated RXR α disruption. WT (+), L2 and L- alleles were identified by PCR on DNA extracted from various organs of K5–Cre–ER^{T(tg/0)}/RXR α ^{L2/+} mice, 12 weeks AFT. **d**, Tamoxifen-induced generation of RXR α null alleles in adult mouse epidermis using K14–Cre–ER^{T2(tg/0)} or K14–Cre–ER^{T2(0/0)} mice (designated (tg/0) and (0/0), respectively). PCR analysis of genomic DNA from epidermis (E) and dermis (D), isolated two weeks after injection of either Tam (0.1 mg) (+) or vehicle (-). Mouse genotypes are indicated and PCR fragments corresponding to RXR α (+), L2 and L- alleles are displayed.

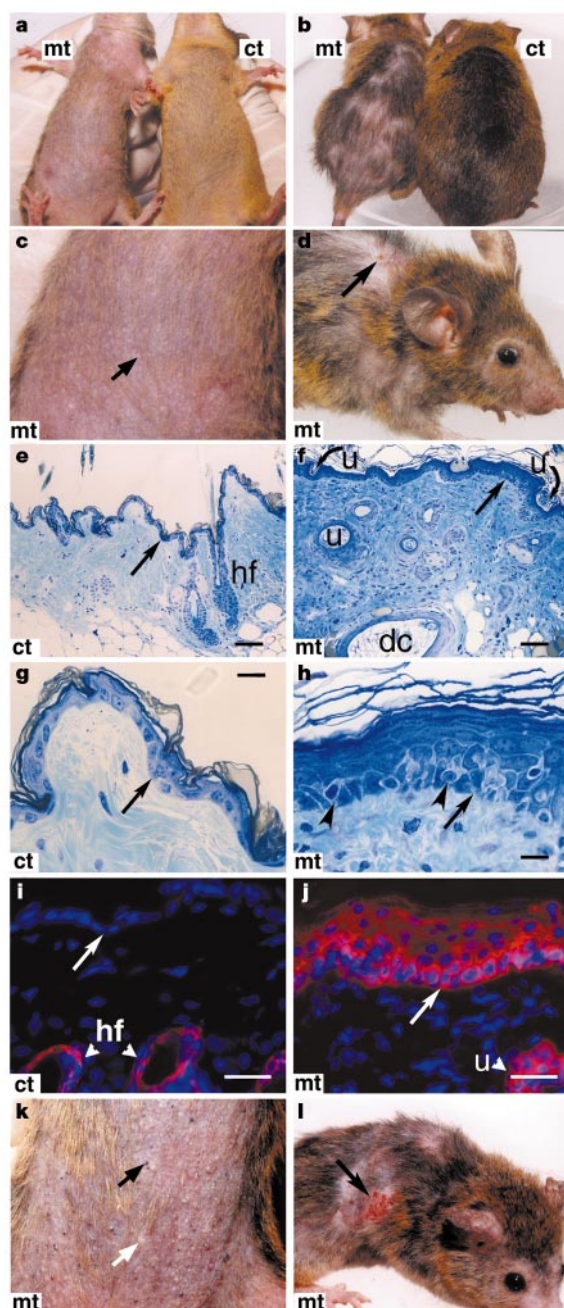


Figure 2 Abnormalities generated by Tam-induced disruption of RXR α in skin of adult mouse mediated by K5-Cre-ER^{T2} and K14-Cre-ER^{T2}. **a, b**, A female K5-Cre-ER^{T2}/RXR α ^{L2/L2} 'mutant' (mt) mouse, and a female K5-Cre-ER^{T2}/RXR α ^{L2/L2} 'control' (ct) mouse, 16 weeks AFT (1 mg Tam per injection). **a**, Ventral view. **b**, Dorsal view. **c**, Higher magnification of the ventral region of the K5-Cre-ER^{T2}/RXR α ^{L2/L2} mouse. Arrow, one of the cysts. **d**, Dorsal view of a female K5-Cre-ER^{T2}/RXR α ^{L2/L2} mouse, 28 weeks AFT. Arrow, minor skin lesion. **e-h**, Histological analysis. 2- μ m sections of ventral skin 16 weeks AFT, taken from 'control' (**e, g**) and 'mutant' (**f, h**) mice. hf, hair follicles; u, utriculi; dc, dermal cysts; arrowheads (**h**), Langerhans cells whose number is increased several-fold in mutant epidermis. **i, j**, Keratin 6 (K6) immunohistochemistry on 'control' (**i**) and 'mutant' (**j**) skin sections (16 weeks AFT). Red, staining of the K6 antibody; cyan, DAPI staining. Arrow (**e-j**), the dermal-epidermal junction. **k, l**, Skin appearance of a female K14-Cre-ER^{T2}/RXR α ^{L2/L2} 'mutant' mouse. **k**, High magnification of the ventral region, 16 weeks after Tam treatment (0.1 mg per injection). White arrow, a cyst; black arrow, a melanosome-containing utriculus. **l**, Dorsal view of the same mutant. Arrow, a skin lesion. Scale bars: **e, f**, 60 μ m; **g, h**, 12 μ m; **i, j**, 25 μ m.

At 16 weeks AFT, histological analysis of ventral and dorsal hairless regions showed hair follicle degeneration, resulting in utriculi and dermal cysts^{20,21} (Fig. 2 compare **e** with **f**; and data not shown). Interfollicular epidermis was hyperplastic with increased incorporation of 5-bromodeoxyuridine (BrdU) and expression of the Ki67 proliferation marker (data not shown). Dermal cellularity was increased and capillaries were dilated (Fig. 2 compare **e, g** with **f, h**; and data not shown) underneath the thickened epidermis, reflecting an inflammatory reaction (data not shown). Keratin 6 (K6), which is usually expressed only in hair follicle outer root sheath (ORS), was also expressed in hyperproliferative interfollicular epidermis (Fig. 2i, j), indicating abnormal keratinocyte terminal differentiation²². All abnormalities were less severe, and/or appeared later in males than in females.

To improve the efficiency of Tam-induced Cre-mediated recombination, we engineered K14-Cre-ER^{T2} transgenic lines. The K14 promoter is selective for the basal layer of stratified squamous epithelia²³, and Cre-ER^{T2} can be induced by a milder Tam treatment (0.1 mg for 5 days)¹⁹. We treated 8–10-week-old K14-Cre-ER^{T2}(tg/0)/RXR α ^{L2/L2} mice with Tam, together with 'control' littermates of genotype K14-Cre-ER^{T2}(tg/0)/RXR α ^{L2/+}, K14-Cre-ER^{T2}(0/0)/RXR α ^{L2/+} and K14-Cre-ER^{T2}(0/0)/RXR α ^{L2/L2}. In two weeks, RXR α ^{L2} alleles were fully converted into RXR α ^{L-/-} alleles in epidermis (Fig. 1d, lanes 1 and 7), but not in dermis (Fig. 1d, lanes 2 and 8) of K14-Cre-ER^{T2}-expressing transgenic mice, demonstrating the higher efficiency of Cre-ER^{T2} for mediating the selective somatic mutation of floxed RXR α in epidermis after Tam treatment. No L2 to L- allele conversion occurred in 'controls' lacking the K14-Cre-ER^{T2} transgene (Fig. 1d, lanes 5, 6, 11 and 12) or without Tam treatment (Fig. 1d, lanes 3, 4, 9 and 10). Moreover, 8 weeks after Tam treatment, only the RXR α ^{L-/-} allele was detected in K14-Cre-ER^{T2}(tg/0)/RXR α ^{L2/L2} mouse epidermis (data not shown), indicating that RXR α was disrupted in most, if not all epidermal stem cells. RXR α disruption also occurred in other epithelia in which the K14 promoter is active²⁴ (for example, tongue, oesophagus and stomach, data not shown). Starting 6 weeks after Tam treatment, K14-Cre-ER^{T2}(tg/0)/RXR α ^{L2/L2} mice exhibited abnormalities similar to those observed in Tam-treated K5-Cre-ER^{T2}(tg/0)/RXR α ^{L2/L2} mice, that is, marked hair loss with visible cysts and focal skin lesions appearing at later stages (Fig. 2k and l; and data not shown). The underlying epidermal and dermal histological abnormalities were also similar to those described above for Tam-treated K5-Cre-ER^{T2}(tg/0)/RXR α ^{L2/L2} mice (data not shown).

Thus, disruption of floxed RXR α in adult epidermis is achieved faster and with lower Tam doses in K14-Cre-ER^{T2} than in K5-Cre-ER^{T2} mice, but the resulting skin abnormalities are similar and, in both cases, more severe in females than in males. Interestingly, these abnormalities are also similar to those exhibited by K14-Cre^(tg/0)/RXR α ^{L2/L2} or K14-Cre^(tg/0)/RXR α ^{L2/-} mice in which floxed RXR α alleles are selectively disrupted in the epidermis during fetal development, leading to RXR α ablation in epidermal keratinocytes and hair follicle ORS (M.L. *et al.*, manuscript in preparation). Indeed, from three weeks of age, these 'constitutive' epidermis-selective RXR α mutants develop progressive alopecia with typical features of degenerated hair follicles, together with utriculi and dermal cysts, which can all be attributed to defects in hair cycles^{20,21}. Furthermore, these mutants also exhibit interfollicular keratinocyte hyperproliferation, as well as abnormal terminal differentiation (with K6 expression) and increased dermal cellularity associated with a skin inflammatory reaction (M.L. *et al.*, manuscript in preparation).

RXR β is expressed at a much lower level than is RXR α in mouse skin, and RXR γ expression is undetectable by polymerase chain reaction after reverse transcription of RNA (RT-PCR; data not shown). Interestingly, we found that the adult skin level of RXR β messenger RNA is several fold higher in males than in females. However, the skin of adult male and female RXR β ^{-/-} mutants

appears normal²⁵ (data not shown) and RXR β mRNA levels did not change upon RXR α ablation (data not shown). Compound mutants were generated to explore a possible functional redundancy between RXRs.

As expected, oil-treated K5-Cre-ER^{T(tg/0)}/RXR α ^{L2/L2}/RXR β ^{-/-} and K5-Cre-ER^{T(tg/0)}/RXR α ^{L2/L2}/RXR β ^{-/-} mice did not exhibit skin abnormalities, but 4 weeks after treatment with Tam they began to lose their hair. Large skin regions were hairless 16–18 weeks AFT, and epidermal flaking on the hairless trunk, chin and ears was much more conspicuous than on single RXR α mutants (compare Fig. 3a; with Fig. 2d; and data not shown). Crusted skin lesions and ulcers lacking epidermis, not seen in RXR α single mutants, were also frequently observed on these RXR α /RXR β double mutants at 14–16 weeks AFT, particularly on the hairless trunk skin, behind the ears and around the mouth (Fig. 3a; and data not shown). But wound repair was not overtly impaired in these mutants when skin biopsies were taken from lesion-free hairless regions (data not shown). Histology of hairless skin showed disappearance of hair follicles and presence of utriculi and dermal cysts (Fig. 3b). The epidermis was highly hyperplastic and hyperkeratinized (compare Fig. 3b, c with Fig. 2e, g; Fig. 2f, h), abnormal K6 expression was observed throughout epidermis (Fig. 3d), and an inflammatory reaction with increased dermal cellularity was also observed (Fig. 3b; and data not shown).

The K5-Cre-ER^{T(tg/0)}/RXR α ^{L2/L2}/RXR β ^{-/-}/RXR γ ^{-/-} triple mutants treated with Tam did not reveal any further role of RXR γ in adult skin (data not shown). Thus, RXR β can partially compensate for a loss of RXR α function. Also, in accordance with the larger amount of RXR β in adult male skin, the functional redundancy was more pronounced in males than in females as RXR α /RXR β double mutant males and females were similarly affected (data not shown) unlike the single mutants (see above).

Collectively, our results demonstrate the effectiveness of Cre-ER^T

recombinases in generating cell-specific temporally controlled targeted somatic mutations in adult mouse tissues. We also show that RXR α , whose knockout is lethal *in utero*^{16,17}, is important postnatally in processes controlling hair cycling and the proliferation and differentiation of epidermal keratinocytes, even though expression of a dominant-negative RXR α mutant in epidermal suprabasal layers has no effect on skin development and maintenance¹⁵.

Our study also reveals a functional redundancy between RXR α and RXR β , although RXR α function is clearly dominant. The molecular events underlying the generation of alopecia and keratinocyte abnormalities in epidermis of mice lacking these receptors are unknown. However, a number of nuclear receptors (for example, RARs, TRs, VDR, PPARs, LXRs) form heterodimers with RXR, and numerous *in vitro* studies using cultured cells and a few *in vivo* targeted-mutagenesis studies^{10–12,26} show that these heterodimers act as signal transducers in different signalling pathways. Interestingly, VDR is also expressed in hair follicle ORS²⁷ and VDR knockout mice develop progressive secondary alopecia, indicating that VDR is important in hair cycling^{2–4}. Alopecia developed by 14-week-old VDR^{-/-} mice appears similar externally to that developed by K5-Cre-ER^{T(tg/0)}/RXR α ^{L2/L2}/RXR β ^{-/-} mice 18 weeks after Tam treatment, although the skin of VDR^{-/-} mice was free of the lesions seen on RXR-ablated epidermis (compare Fig. 3a with e). At the histological level, similar utriculi and dermal cysts were observed (compare Fig. 3b with f), but no keratinocyte hyperproliferation was observed in the epidermis of VDR^{-/-} mice and keratinocyte differentiation was normal (as revealed by K6 expression (compare Fig. 3c and d with g and h; and data not shown). Thus, alopecia generated by selective RXR ablation in adult mouse keratinocytes may reflect a major role of RXR/VDR heterodimers in hair cycling. Further keratinocyte-specific targeted somatic mutagenesis is necessary to investigate whether other signalling pathways involving nuclear receptors that heterodimerize with RXRs (notably RARs¹, LXRs⁷ and PPARs^{8,28}) are implicated in the generation of the other skin abnormalities resulting from keratinocyte-selective RXR ablation. □

Methods

Transgenic lines

RXR α ^{L2/L2}, VDR^{-/-} and K5-Cre-ER^T mouse lines have been described^{2,16,19} and RXR α ^{L2/+} mice will be described (M.L. *et al.*, manuscript in preparation). The K14-Cre-ER^{T2} transgene was constructed by replacing the K5 promoter region of pK5-Cre-ER^{T2} (ref. 19) by the 2-kilobase (kb) human keratin K14 promoter/enhancer *SalI* DNA fragment²³, isolated from pHR2 (a gift from S. Werner), and transgenic mice were generated¹⁹.

Genotyping of RXR α alleles

Genomic DNA was isolated from tissues as described¹⁹. Epidermis was separated from dermis after treating tail skin with dispase (4 mg ml⁻¹ in PBS, Gibco-BRL) for 1–2 h at room temperature. RXR α genotyping was performed by PCR. Primers: ZO243 (5'-TCCTTCACCAAGCACATCTG-3', in exon 3) and ZO244 (5'-TGCAGCCCTCACAAC TGTAT-3', in exon 4) for L2 and (+) alleles (700- and 650-base pair (bp) fragment, respectively); ZO243 and UD196 (5'-CAACCTGGACTTGTCACTTAG-3' in the intron between exons 4 and 5) for L- allele (400-bp fragment); ZO243 and RU178 (5'-ATGTTTCATAGTTGGATATC-3', in *neo* cassette) for (-) allele (500-bp fragment). For Southern blot analysis, genomic DNA was digested with *Bam*HI and probed with probe X4 (3-kb *Bam*HI–*Xba*I fragment of RXR α gene)²⁹.

Tamoxifen treatment

Tam (Sigma) solutions have been described¹⁸. Tam (1 mg in 100 μ l sunflower oil) was injected intraperitoneally into K5-Cre-ER^T transgenic mice for five consecutive days, and again for three consecutive days, two, four and six weeks later. K14-Cre-ER^{T2} transgenic mice were injected intraperitoneally with 0.1 mg Tam (in 100 μ l sunflower oil) for five consecutive days.

Histological analysis

Skin biopsies of age- and sex-matched animals were taken from similar body sites. Skin samples were fixed in glutaraldehyde (2.5% in 0.1 M cacodylate buffer pH 7.2) overnight at 4 °C, and post-fixed with 1% osmium tetroxide in cacodylate buffer for 1 h at 4 °C. Tissues were dehydrated with graded concentrations of alcohol and embedded in Epon 812. Semi-thin sections (2 μ m) were stained with toluidine blue.

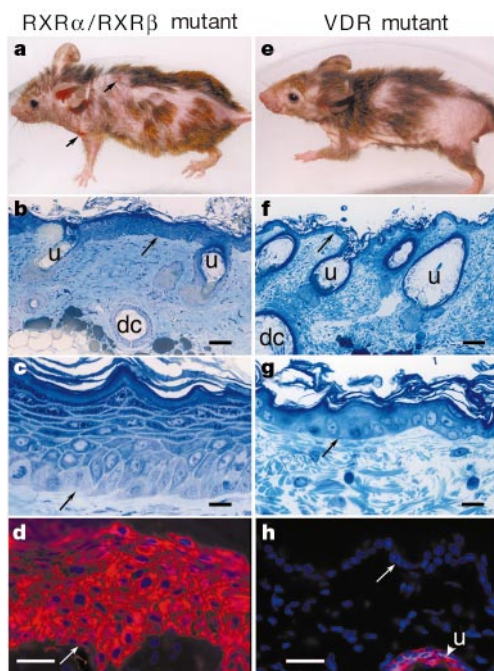


Figure 3 Comparison of skin abnormalities exhibited by a Tam-treated K5-Cre-ER^{T(tg/10)} RXR α ^{L2/L2}/RXR β ^{-/-} mouse and a VDR-null mouse. The figure corresponds to a K5-Cre-ER^{T(tg/0)}/RXR α ^{L2/L2}/RXR β ^{-/-} mouse, 18 weeks AFT (1 mg Tam per injection) (a–d) and to a 14-week-old VDR^{-/-} mouse (e–h). **b, c, f, g**, Histological analysis of 2- μ m dorsal skin sections. **d, h**, Keratin K6 immunohistochemistry on skin sections. u, utriculi; dc, dermal cysts. Arrows: **a**, skin lesions; **b–h**, the dermal–epidermal junction. Scale bar: **b, f**, 60 μ m; **c, g**, 12 μ m; **d, h**, 25 μ m.

Immunohistochemistry

After fixation in 2% paraformaldehyde, 10-µm frozen sections were blocked in 5% normal goat serum (Vector), incubated with rabbit polyclonal anti-MK6 (Babco). After washing in PBS containing 0.1% Tween 20, sections were incubated with CY3-conjugated donkey anti-rabbit IgG antibody (Jackson ImmunoResearch), and mounted with Vectashield medium (Vector) containing DAPI (4,6-diamidino-2-phenylindole; Boehringer Mannheim)³⁰.

Received 22 May; accepted 28 July 2000.

1. Fisher, G. J. & Voorhees, J. J. Molecular mechanisms of retinoid actions in skin. *FASEB J.* **10**, 1002–1013 (1996).
2. Yoshizawa, T. *et al.* Mice lacking the vitamin D receptor exhibit impaired bone formation, uterine hypoplasia and growth retardation after weaning. *Nature Genet.* **16**, 391–396 (1997).
3. Li, Y. C. *et al.* Targeted ablation of the vitamin D receptor: an animal model of vitamin D-dependent rickets type II with alopecia. *Proc. Natl Acad. Sci. USA* **94**, 9831–9835 (1997).
4. Li, Y. C. *et al.* Normalization of mineral ion homeostasis by dietary means prevents hyperparathyroidism, rickets, and osteomalacia, but not alopecia in vitamin D receptor-ablated mice. *Endocrinology* **139**, 4391–4396 (1998).
5. Kömives, L. G. *et al.* Ligands and activators of nuclear hormone receptors regulate epidermal differentiation during fetal rat skin development. *J. Invest. Dermatol.* **111**, 429–433 (1998).
6. Hanley, K. *et al.* Activators of the nuclear hormone receptors PPAR α and FXR accelerate the development of the fetal epidermal permeability barrier. *J. Clin. Invest.* **100**, 705–712 (1997).
7. Hanley, K. *et al.* Oxysterols induce differentiation in human keratinocytes and increase Ap-1-dependent involucrin transcription. *J. Invest. Dermatol.* **114**, 545–553 (2000).
8. Hanley, K. *et al.* Farnesol stimulates differentiation in epidermal keratinocytes via PPAR α . *J. Biol. Chem.* **275**, 11484–11491 (2000).
9. Mangelsdorf, D. J. *et al.* The nuclear receptor superfamily: the second decade. *Cell* **83**, 835–839 (1995).
10. Chambon, P. A decade of molecular biology of retinoic acid receptors. *FASEB J.* **10**, 940–954 (1996).
11. Kastner, P., Mark, M. & Chambon, P. Nonsteroid nuclear receptors: what are genetic studies telling us about their role in real life? *Cell* **83**, 859–869 (1995).
12. Mascres, B. *et al.* The RXR α ligand-dependent activation function 2 (AF-2) is important for mouse development. *Development* **125**, 4691–4707 (1998).
13. Saitou, M. *et al.* Inhibition of skin development by targeted expression of a dominant-negative retinoic acid receptor. *Nature* **374**, 159–162 (1995).
14. Imakado, S. *et al.* Targeting expression of a dominant-negative retinoic acid receptor mutant in the epidermis of transgenic mice results in loss of barrier function. *Genes Dev.* **9**, 317–329 (1995).
15. Feng, X. *et al.* Suprabasal expression of a dominant-negative RXR α mutant in transgenic mouse epidermis impairs regulation of gene transcription and basal keratinocyte proliferation by RAR-selective retinoids. *Genes Dev.* **11**, 59–71 (1997).
16. Kastner, P. *et al.* Genetic analysis of RXR α developmental function: convergence of RXR and RAR signalling pathways in heart and eye morphogenesis. *Cell* **78**, 987–1003 (1994).
17. Sucov, H. M. *et al.* RXR alpha mutant mice establish a genetic basis for vitamin A signaling in heart morphogenesis. *Genes Dev.* **8**, 1007–1018 (1994).
18. Metzger, D. & Chambon, P. Site- and time-specific gene targeting in the mouse. *Methods* (in the press).
19. Indra, A. K. *et al.* Temporally-controlled site-specific mutagenesis in the basal layer of the epidermis: comparison of the recombinase activity of the tamoxifen-inducible Cre-ER^T and Cre-ER^{T2} recombinases. *Nucleic Acids Res.* **27**, 4324–4327 (1999).
20. Sundberg, J. P. & King, L. E. Mouse models for the study of human hair loss. *Dermatol. Clin.* **14**, 619–632 (1996).
21. Panteleyev, A. A., Paus, R., Ahmad, W., Sundberg, J. P. & Christiano, A. M. Molecular and functional aspects of the hairless (hr) gene in laboratory rodents and humans. *Exp. Dermatol.* **7**, 249–267 (1998).
22. Porter, R. M., Reichelt, J., Lunny, D. P., Magin, T. M. & Lane, B. The relationship between hyperproliferation and epidermal thickening in a mouse model for BCIE. *J. Invest. Dermatol.* **110**, 951–957 (1998).
23. Vassar, R., Rosenberg, M., Ross, S., Tyner, A. & Fuchs, E. Tissue-specific and differentiation-specific expression of a human K14 keratin gene in transgenic mice. *Proc. Natl Acad. Sci. USA* **86**, 1563–1567 (1989).
24. Wang, X., Zinkel, S., Polonsky, K. & Fuchs, E. Transgenic studies with a keratin promoter-driven growth hormone transgene: prospects for gene therapy. *Proc. Natl Acad. Sci. USA* **94**, 219–226 (1997).
25. Kastner, P. *et al.* Abnormal spermatogenesis in RXR β mutant mice. *Genes Dev.* **10**, 80–92 (1996).
26. Wendling, O., Chambon, P. & Mark, M. Retinoid X receptors are essential for early mouse development and placental development. *Proc. Natl Acad. Sci. USA* **96**, 547–551 (1999).
27. Reichrath, J. *et al.* Hair follicle expression of 1,25-dihydroxyvitamin D3 receptors during the murine hair cycle. *Br. J. Dermatol.* **131**, 477–482 (1994).
28. Braissant, O., Foulle, E., Scotto, C., Dauca, M. & Wahli, W. Differential expression of peroxisome proliferator-activated receptors (PPARs): tissue distribution of PPAR- α , - β , and - γ in the adult rat. *Endocrinology* **137**, 354–366 (1996).
29. Metzger, D. *et al.* Conditional site-specific recombination in mammalian cells using a ligand-dependent chimeric Cre recombinase. *Proc. Natl Acad. Sci. USA* **92**, 6991–6995 (1995).
30. Brocard, J. *et al.* Spatio-temporally controlled site-specific somatic mutagenesis in the mouse. *Proc. Natl Acad. Sci. USA* **94**, 14559–14563 (1997).

Acknowledgements

We thank S. Werner for the human K14 promoter; H. Chiba and P. Kastner for RXR α ^{L2/+}, RXR α ^{+/-} and RXR β ^{+/-} mice; J. M. Bornert, S. Bronner, N. Chartoire, M. Duval, C. Gérard, R. Lorentz and J. L. Vonesch for technical help; M. LeMeur, R. Matyas and the animal facility staff for animal care; M. Mark for histological analysis; the secretariat for typing the manuscript and the illustration staff for preparing the figures; and all the members of the laboratory for helpful discussions. This work was supported by funds from the Centre National de la Recherche Scientifique, the Institut National de la Santé et de la Recherche Médicale, the Collège de France, the Hôpital Universitaire de Strasbourg, the Association pour la Recherche sur le Cancer, the Fondation pour la Recherche Médicale, the Human Frontier Science Program, the Ministère de l'Éducation Nationale de la Recherche et de la

Technologie and the European Economic Community. M.L. was supported by fellowships from the Association pour la Recherche sur le Cancer and the Fondation pour la Recherche Médicale, A.K.I. by a fellowship from the Université Louis Pasteur (Strasbourg), and J.B. and X.W. by fellowships from the Ministère de l'Éducation Nationale, de la Recherche et de la Technologie and from the Fondation pour la Recherche Médicale.

Correspondence and requests for materials should be addressed to P.C. (e-mail: chambon@igbmc.u-strasbg.fr).

Memory B-cell persistence is independent of persisting immunizing antigen

Mitsuo Maruyama*, Kong-Peng Lam*† & Klaus Rajewsky*

* Institute for Genetics, University of Cologne, Weyertal 121, D-50931 Cologne, Germany

† Institute of Molecular and Cell Biology, The National University of Singapore, 30 Medical Drive, Singapore 117609

Immunological memory in the antibody system is generated in T-cell-dependent responses and carried by long-lived memory B cells that recognize antigen by high-affinity antibodies^{1,2}. But it remains controversial¹ whether these B cells represent true 'memory' cells (that is, their maintenance is independent of the immunizing antigen), or whether they are a product of a chronic immune response driven by the immunizing antigen, which can be retained in the organism for extended time periods on the surface of specialized antigen-presenting cells (follicular dendritic cells)³. Cell transfer experiments provided evidence in favour of a role of the immunizing antigen^{4,5}; however, analysis of memory cells in intact animals, which showed that these cells are mostly resting⁶ and can persist in the absence of detectable T-cell help⁷ or follicular dendritic cells⁸, argued against it. Here we show, by using a genetic switch mediated by Cre recombinase, that memory B cells switching their antibody specificity away from the immunizing antigen are indeed maintained in the animal over long periods of time, similar to cells retaining their original antigen-binding specificity.

The main difficulty in the analysis of the antigen-dependence of memory B-cell persistence lies in the elimination of residual antigen from the cellular environment, which is difficult to achieve beyond doubt. We therefore looked for an alternative approach, namely to use induced targeted mutagenesis *in vivo* to equip memory B cells induced by and specific for a particular antigen with an antigen receptor (BCR) that does not bind the inducing antigen; and to study the persistence of the mutant cells in comparison with others that do not switch BCR specificity.

The genetic switch had to be irreversible, in contrast to the system of BCR specificity switching that we previously developed⁹. We therefore generated a mouse strain carrying two distinct, mutant alleles of the immunoglobulin (Ig) heavy (H) chain (IgH) locus. One of these alleles was already available and carried a VDJ segment in its physiological position, flanked by loxP sites in the same orientation¹⁰. This IgH allele, designated VNP here, encodes an H chain that, together with λ 1 light (L) chains, forms an antibody with specificity for the hapten 4-hydroxy-3-nitro-phenylacetyl (NP).

The second IgH allele (Fig. 1) contains a VDJ segment encoding, again in the context of λ 1 L chains, an antibody specific for the fluorescent protein phycoerythrin (PE). This VDJ segment, designated VDJPE, originated from a hybridoma isolated from a PE-immunized mouse, selected for reactivity with PE but not NP (see

Methods). We chose PE specificity because PE-specific memory B cells can be easily monitored by flow cytometry^{6,11}. VDJPE was flanked by mutant *loxP* sites¹² in opposite orientations, allowing Cre-mediated inversion between them (Fig. 1b) but not transchromosomal recombination with wild-type *loxP* sites on the *VNP* allele (see below). Thus, Cre recombinase can change an IgH allele carrying an inverted VDJPE gene segment (designated *VPEinv*) to an IgH allele allowing expression of an H chain that, together with $\lambda 1$ L chains, binds PE but not NP (see Fig. 2a).

The mouse that we used for the BCR specificity switch experiments (Fig. 2a) carries the *VNP* and *VPEinv* IgH alleles, together with a *cre* transgene (*Mx-cre*) that expresses Cre on induction with type I interferon (IFN)¹³. Essentially all B cells generated in this mouse express the *VNP* allele¹⁰, and hence NP-binding antibodies if $\lambda 1$ L chains are co-expressed. This is expected for about 4% of the B cells, which is what we observed (Fig. 2b).

On IFN injection, the percentage of NP-binding cells decreases and PE-binding cells appear, reaching levels of about 3% at day 10 (Fig. 2b). These results are in agreement with Southern blot analyses of liver and splenic B cells from such animals, allowing a quantitative determination of *VNP* gene segment deletion (generating the *VNPΔ* allele) and *VPEinv* gene inversion (generating the *VPE* allele; Fig. 1a, c). For liver cells, a single IFN injection caused nearly a 70% deletion of the *VNP* gene segment, and 35–40% inversion of *VPEinv*. (Note that recombination generates an equilibrium between the two orientations.) Similar results are obtained for splenic B cells on day 3 after IFN treatment. On day 10, however, the fraction of the *VNPΔ* allele drops in these cells to roughly 40%, and that of the *VPE* allele rises to over 60%. This result is expected, because B (but not liver) cells that have undergone deletion of the *VNP* allele and not inverted *VPEinv* lose BCR expression, resulting in apoptotic death after 1–2 weeks¹⁰.

Assuming that Cre-mediated recombination is induced by IFN roughly equally in liver cells and lymphocytes¹³, we can calculate from the Southern blot data on liver cells (Fig. 1c) the expected loss of B cells and the fractions of *VNPΔ* and *VPE* alleles in the remaining population, and compare these with experimental values. Doing this, the calculated fraction of surviving B cells is 57%, which corresponds well with the 54% observed on day 10 after IFN injection, based on the analysis of three experimental and three control mice; on day 3 after IFN injection, we measured 78% B cells compared with B cells in the spleens of *VNP/VPEinv* controls. On the basis of the cellular yield on day 10, we calculate fractions of 44% for *VNPΔ* and 65% for *VPE* in those cells, in agreement with the Southern blot data (Fig. 1c). Thus, for the splenic B-cell population, our system allows the *in vivo* generation of a sizeable fraction of cells that have switched their BCR specificity from NP to PE on IFN-mediated induction of Cre recombinase.

We aimed for the induction NP-specific memory cells, followed by the induction of a switch of BCR specificity to PE in some of these cells and a subsequent comparative analysis of the *in vivo* maintenance of the NP- and PE-specific cells over time. We immunized *VNP/VPEinv* mice carrying the *Mx-cre* transgene with a NP/chicken γ -globulin (CG) conjugate in adjuvant (Fig. 3a). Six to eight weeks later, after the establishment of NP-specific memory, the mice were given a single dose of 1×10^6 units IFN or phosphate-buffered saline (PBS) for control. At various times thereafter, we isolated and quantified splenic NP- and PE-binding memory B cells.

Because the secondary response to NP/CG consists mainly of IgG1 antibodies, implying that most memory cells express IgG1 antibodies in their BCRs¹⁴, we enriched IgG1-expressing B cells (about 1% of the total B-cell population) by magnetic cell sorting (MACS) and subsequently analysed them for NP- and PE-binding by flow cytometry. The purification procedure did not detectably affect the distribution of these specificities in the IgG1⁺ cell population (data not shown). Typical results obtained from experimental and control animals are shown in Fig. 3b. In NP-immunized

VNP/VPEinv mice, roughly one-third of the IgG1⁺ cells bind NP (left panels). The presence of these cells depends upon immunization (lower right panel), and more than 90% of the NP⁺ cells express $\lambda 1$ L chains (data not shown). In addition to these cells, small fractions of PE⁺ cells can be detected after IFN injection (upper panels). These cells are 6–7 times less abundant than the NP⁺PE⁺ cells, and 20–30% of them bind NP in addition to PE. PE⁺ cells are barely detectable after PBS injection (lower left panels), their

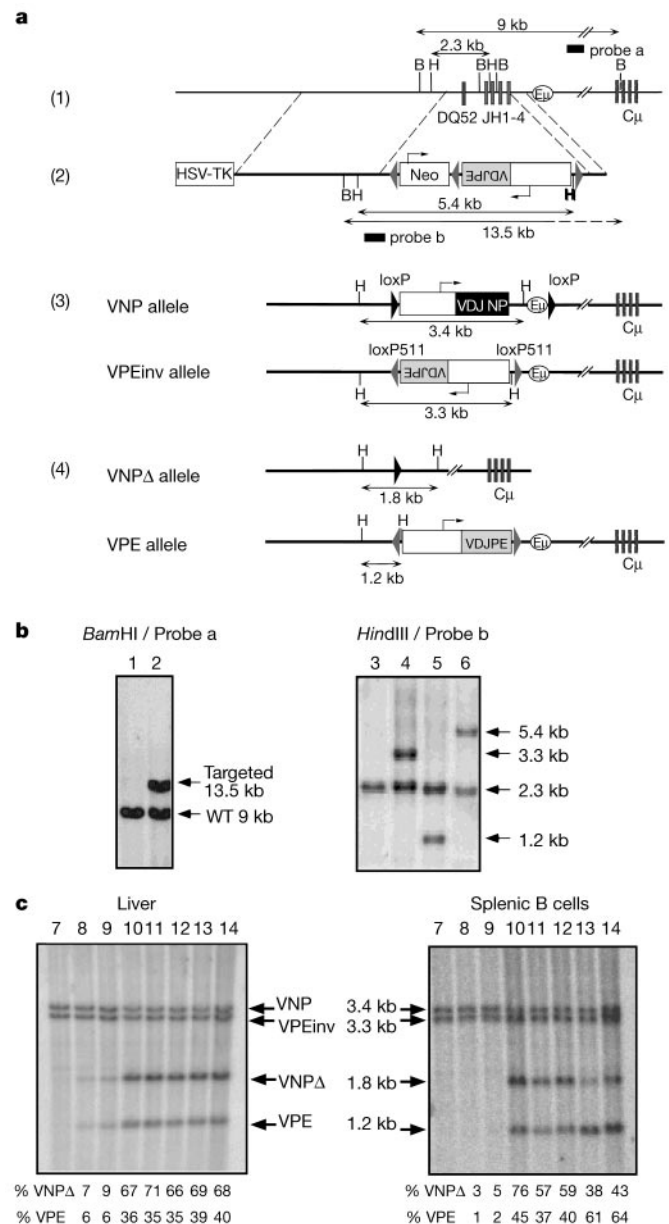


Figure 1 Generation of the *VPEinv* allele and efficiency of Cre-mediated gene modification in *VNP/VPEinv*, *Mx-cre* mice. **a**, Targeting strategy (see Methods) and partial restriction map of the *IgH* loci in *VNP/VPEinv*, *Mx-cre* mice. **b**, Southern blot of E14 and targeted ES-cell DNA digested with *Bam*HI (B), and homologous recombinant ES clone DNA transfected with *pC-cre* digested with *Hind*III (H). Hybridized probes, fragments and their sizes are shown. Lanes 1 and 3, parental E14; lanes 2 and 6, homologous recombinant; lane 4, deletion of the *neo*^r gene only; lane 5, deletion of the *neo*^r gene in addition to inversion of *VPEinv*. **c**, Southern blot of genomic DNA from liver (left) and splenic B cells (right) of *VNP/VPEinv* (lane 7) and *VNP/VPEinv*, *Mx-cre* mice (lanes 8–14) 3 days (lanes 8, 10, 11, 12) or 10 days (lanes 9, 13, 14) after a single dose of PBS (lanes 8, 9) or IFN injection (lanes 10–14), digested with *Hind*III and hybridized with probe b. Splenic B cells were purified by using anti-CD19 MACS beads. Percentages deletion and inversion are indicated under each lane.

presence depends on NP immunization and they do not appear spontaneously in IgG1⁺ cells from *VPE/VPE^{inv}* mice (lower right panel). Although the percentage of NP⁺PE⁺ cells seems to decline over time, the PE⁺ cells are maintained in the animals at constant levels over 105 days after IFN injection, like NP⁺ cells in IFN- or PBS-injected animals (Fig. 3c).

The low efficiency at which of memory (as compared with naive) cells with 'switched' BCR specificity are generated in our system is expected, as the inverted *VPE* alleles carried along by the cells in the NP-driven germinal centre (GC) reaction (in which memory cells are generated; see below) accumulated unselected mutations. Only mutants still able to produce a PE-binding antibody and also having undergone class switch recombination to IgG1 on both IgH alleles (which happens in only part of the cells¹⁵) would later be detectable as 'switched' PE-binding cells. A molecular and functional characterization of the 'switched' memory cells is, however, clearly required to prove their identity.

For this purpose, we genotyped PE⁺NP⁻ and PE⁺NP⁺ memory cells by single-cell amplification of the *VPE* and *VNP* alleles in the *IgH* loci and of *Vλ1/λ1* and *Vκ/κ* genomic rearrangements. Despite the limited efficiency of single-cell genomic amplification, the results strongly indicated that, for the majority, the two classes of cells have the expected genotypes: both have undergone inversion of *VPE^{inv}* to generate a (productive) *VPE* allele (detected in 6/7 and 4/7 informative cells, respectively), and in most cells (6/7 and 4/7, respectively) *Vλ1/λ1* rearrangements are detected. In contrast, the *VNP* allele is absent in most PE⁺NP⁻ cells (5/7), whereas it is present in the PE⁺NP⁺ cells (7/7). (The two PE⁺NP⁻ cells in which a signal for the *VNP* allele was detected may carry the *VNP* gene segment on

circular DNA excised by Cre recombinase; data not shown.)

A second molecular approach uses the fact that the switched memory cells should have passed through the GC reaction as NP⁺ cells and undergone somatic hypermutation and selection for high-affinity NP binders^{16,17}. In this phase, the cells carried along the *VPE^{inv}* allele, which, as it possesses all the *cis* regulatory elements controlling somatic hypermutation¹⁸, should have acquired somatic mutations in an unselected way. In addition, as we later isolated memory cells as IgG1-expressing cells, the cells should have undergone class switch recombination¹⁵ to Cγ1 on both IgH alleles.

To address these matters, we amplified *VPE^{inv}* from the genomic DNA of single, NP⁺memory cells and used polymerase chain reaction with reverse transcription (RT-PCR) to amplify *VPE-γ1* transcripts from PE⁺memory cells. Most of both PE⁺NP⁻ and PE⁺NP⁺ memory cells (13/16 in both cases; data not shown) allowed the amplification of a *VPE-γ1* transcript, showing that these cells had indeed undergone class switch recombination on the *VPE* allele. We then sequenced and compared the PCR products from the three classes of cell.

Between 55 and 64% of the sequences carried somatic point mutations with an average mutation frequency of 4.5–5% (data not shown). The pattern of somatic mutation of *VPE^{inv}* in NP⁺, and of *VPE* in PE⁺NP⁺IgG1⁺ cells is shown in Fig. 4 (A similar pattern was found for *VPE* in the PE⁺NP⁺IgG1⁺ cells; data not shown). In all cases, the mutations spread over the entire V region and are strongly biased towards the well-characterized 'hot spots' of somatic hypermutation¹⁸.

Significantly, in 2 of the 9 sequences obtained for *VPE^{inv}* from NP⁺ memory cells a stop codon was generated by a point mutation.

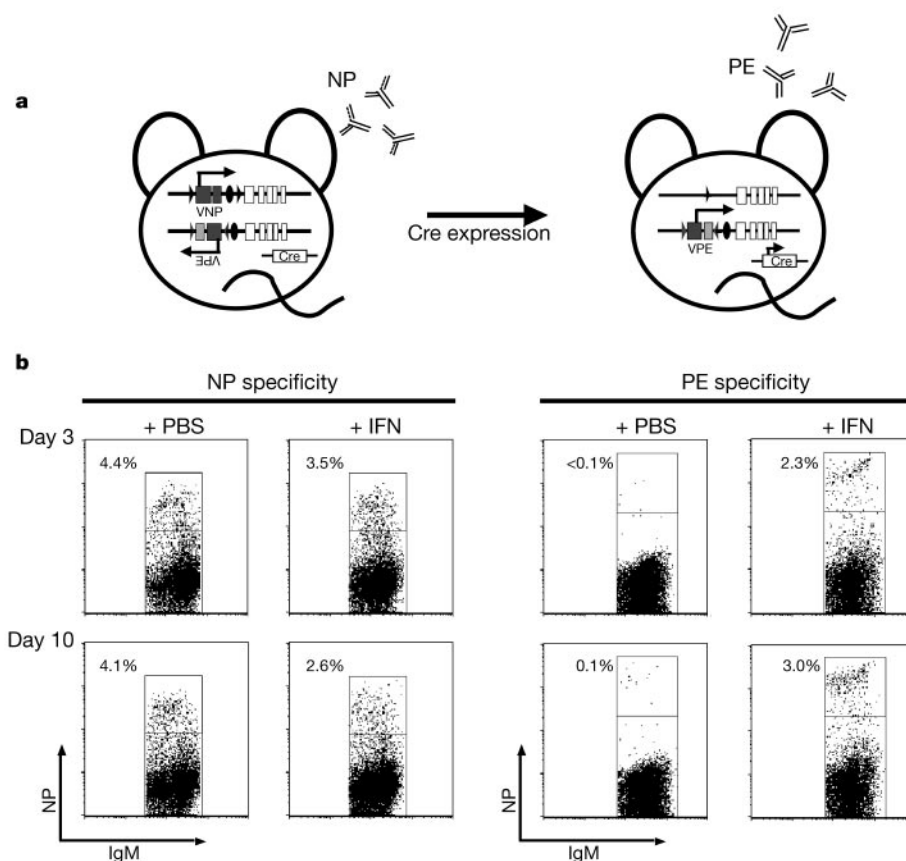


Figure 2 Induced change of antigen specificity on mature B cells *in vivo*. **a**, Experimental strategy. **b**, Splenic B cells from adult *VNP/VPE^{inv}, Mx-cre* mice 3 (top panels) and 10 days (bottom panels) after a single injection of IFN or PBS were analysed by flow cytometry. Cells were treated with NP14-BSA (left) or PE (right) as antigen and stained

with FITC-anti-IgM antibody, followed by APC-S43 for NP-binding cells. In each group, left panels represent control stainings after PBS injection. IgM⁺ cells were gated, and the percentages of total IgM⁺ cells are shown.

Overall, the data are compatible with the idea that the VPE alleles expressed by the 'switched' PE-binding memory cells isolated from NP-immunized mice acquired their somatic mutations as 'passenger' transgenes in the course of an NP-driven GC reaction.

The PE-binding, IgG1-expressing cells appearing in NP-immunized VNP/VPE^{inv} mice on IFN treatment were also analysed with respect to their antigen responsiveness and their resting state. PE-specific memory cells respond vigorously to antigenic stimulation *in vitro*, in the presence of PE-specific T-helper cells^{6,11}. This can be

reproduced for the PE⁺IgG1⁺ 'switched' memory cells generated in our experimental system and isolated on day 10 after IFN treatment (Table 1), indicating that the switch of BCR specificity does not interfere with the functional properties of the cell. Significantly, the PE-specific antibodies produced by the 'switched' memory cells did not crossreact with NP, demonstrating that somatic mutation had not led to a broadened crossreactivity and the cells were indeed unable to 'see' the antigen by which they had originally been induced.

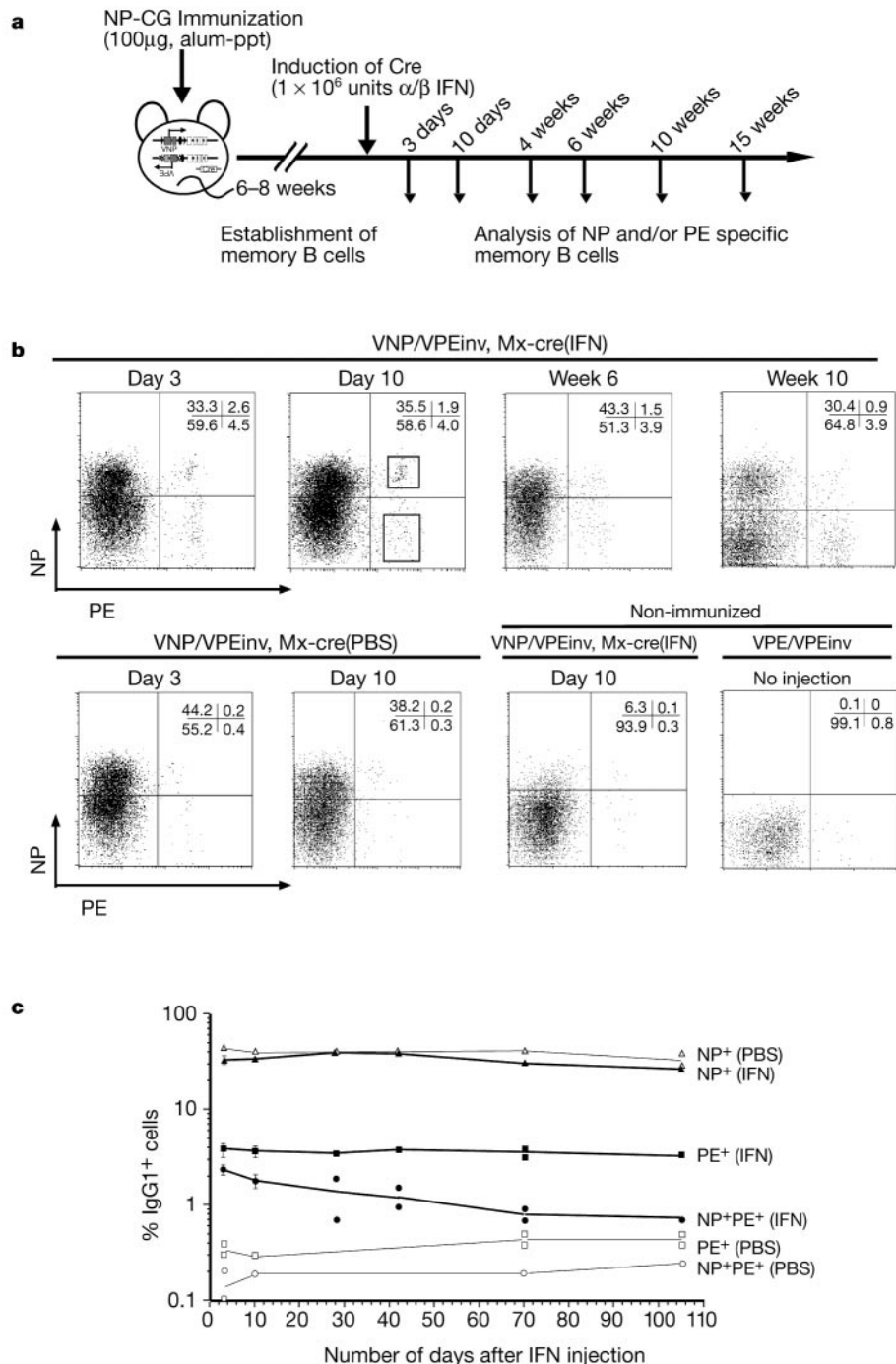


Figure 3 IFN-induced change of antigen specificity on memory B cells *in vivo*.

a, Experimental strategy. **b**, IgG1⁺ splenic B cells from NP/CG-primed VNP/VPE^{inv}, Mx-cre mice isolated at various days after IFN or PBS treatment were stained for NP and PE binding. Controls were non-immunized VNP/VPE^{inv}, Mx-cre and VPE/VPE^{inv} mice. Numbers in quadrants indicate percentages of total IgG1⁺ cells. Boxes indicate gates from

which single cells were sorted. **c**, Persistence of NP/PE-specific IgG1⁺ splenic B cells after IFN injection. Dots represent mean of percentages of IgG1⁺ cells determined for 2–4 individual mice, as in **b**. Thick and thin lines represent the data obtained from IFN- and PBS-injected VNP/VPE^{inv}, Mx-cre mice, respectively.

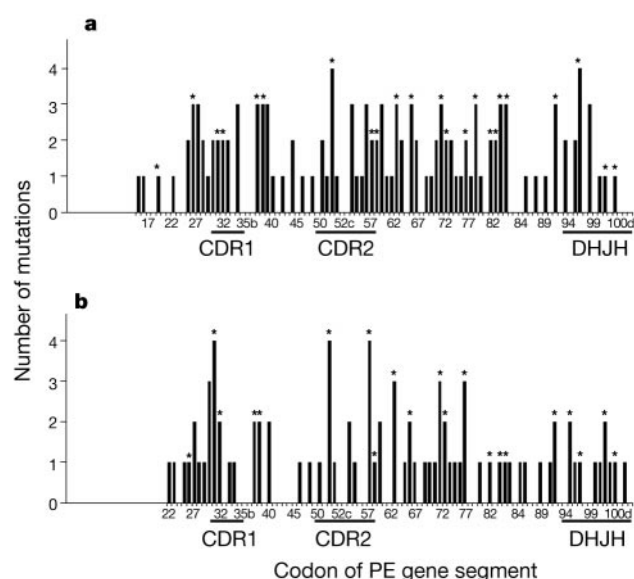


Figure 4 Pattern of somatic mutation in the *VPE* gene of memory B cells. Numbers of mutations per codon sequenced are given. **a**, Inverted *VPE* sequences from individual NP⁺PE[−]IgG1⁺ memory B cells isolated from *VNP/VPEinv* mice primed with NP/CG 8 weeks earlier. **b**, Expressed *VPE* sequences from individual PE⁺NP[−]IgG1⁺ memory B cells on day 10 after IFN injection of NP/CG-primed *VNP/VPEinv*, *Mx-cre* mice. Asterisks indicate positions corresponding to intrinsic mutational hotspots¹⁸, considering both DNA strands. Data are from 9 (**a**) and 7 (**b**) mutated sequences.

Finally, we tested whether the PE-binding IgG1⁺ cells appearing after IFN treatment were in a resting state as they should be, given that memory cells are mostly in a resting state after their generation in the GC reaction⁶. For this purpose, the mice that had been immunized with NP/CG 8 weeks earlier were given 5-bromodeoxyuridine (BrdU) in their drinking water for 3 days, starting either on day 2 or day 5 after IFN or PBS injection. Subsequently, total IgG1⁺ cells and their PE-binding subset were analysed for BrdU incorporation into their DNA, using B-cell progenitors from the bone marrow as high controls. BrdU incorporation into both fractions of memory cells was minimal, contrasting with massive incorporation into the B-cell progenitors (Fig. 5).

Overall, four lines of evidence indicate that the PE-binding IgG1⁺ cells arising in the NP-immunized mutant mice on IFN treatment directly derive from NP-specific memory cells. First, most cells carry heavily mutated *VPE* genes and both the pattern and load of mutation are similar to those of the 'passenger' *VPEinv* alleles in their NP-binding counterparts (Fig. 4). Second, the PE-binding cells appear shortly after IFN injection, at the time expected from the analysis of Cre-mediated recombination (Fig. 1c). Their appearance is dependent on NP/CG immunization, and they do not appear spontaneously in mice expressing *VPE* from the beginning in all B cells (Fig. 3b). Third, the cells do not undergo cell division after IFN treatment but are in a resting state as expected for memory cells after their generation in the GC reaction (Fig. 5; conversely, IFN by itself does also not appear to convert short-lived into long-lived cells, as tested for naive B cells in a system of inducible RAG gene inactivation (Z.-Y. Hao and K.R., unpublished data)). Fourth, the cells share with classical IgG1⁺ memory cells the capacity to produce a vigorous IgG1 response *in vitro* in response to antigenic challenge in the presence of T-helper cells (Table 1).

We observe two types of 'switched' PE-binding memory cells after IFN treatment: a major fraction of cells binding PE but not NP (and having deleted *VNP* and inverted *VPEinv*) and a minor fraction of cells binding both PE and NP. Most of the latter cells have retained the *VNP* allele (data not shown), arguing against the possibility that

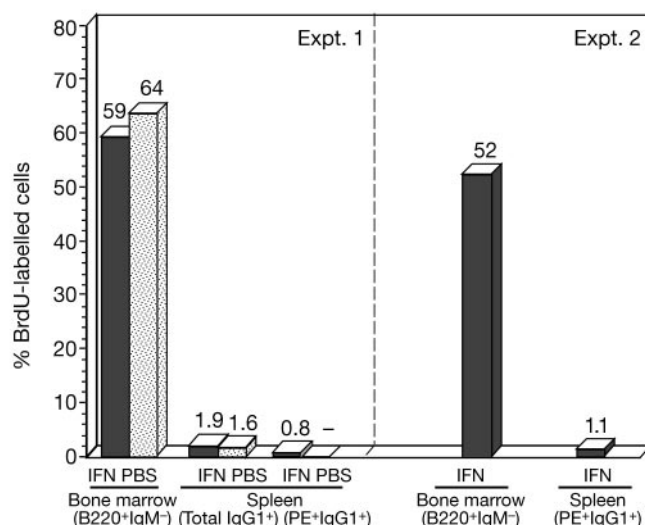


Figure 5 Proliferative activity of 'switched' PE-binding memory B cells. B220⁺IgM[−] bone marrow B cells sorted as positive controls. Numbers above columns represent the percentages of BrdU-labelled cells. BrdU labelling began 2 and 5 days after IFN or PBS injection in experiments 1 and 2, respectively, and lasted for 3 days.

their slow decline over time is due to a slow turnover of surface immunoglobulin on a fraction of memory cells. Although the half life of the double-producing cells appears to be in the order of 50 days, the frequencies of NP⁺PE[−] and PE⁺NP[−] IgG1-expressing memory cells remain essentially constant from day 3 until day 105 after IFN treatment (Fig. 3c). Thus, memory B cells that have switched BCR specificity away from the antigen by which they had been selected in the GC reaction persist in the animal like their counterparts that have not undergone the specificity switch, for at least 105 days. Double-producing memory cells (which have no physiological counterpart as normal B cells express a single antibody specificity) are also long-lived, although their life span seems to be somewhat reduced, for unknown reasons.

The PE⁺ 'switched' memory cells are clearly unable to see the NP hapten by which they had originally been selected in the animal with any appreciable affinity: they not only fail to bind detectably NP-protein conjugate in a flow cytometric assay (in which arguably there might be competition between PE and NP), but they also produce antibodies upon secondary stimulation that fail to cross-react with NP in a sensitive plate binding assay that also detects antibodies of low affinity (Table 1).

Can we exclude that the PE-binding memory cells that appear and persist after IFN treatment recognize some self-antigen(s) with low affinity? Although there is evidence that the *in vivo* maintenance of mature B cells depends on BCR expression¹⁰, it is unresolved whether this is due to a B-cell-autonomous 'maintenance' signal or whether it reflects low-affinity interactions of the BCR with (self-) antigens in the environment. As we isolated the *VPE* allele used in our system from a PE-specific mature B cell from a PE-immunized mouse, it is therefore quite possible that the corresponding antibody possesses some low-level autoreactivity. If this antibody were able to interact with a self-antigen with even moderate affinity, however, it would be counterselected by mechanisms of immunological tolerance in immature as well as mature B cells^{9,19} which is clearly not the case (Figs 2b and 3; and M.M. and K.R., unpublished data on the undisturbed development of PE-specific cells in mice expressing *VPE* in all B cells; a possible enhanced autoreactivity of some of the PE⁺ memory cells that express somatically mutated *VPE* and *VN1* alleles cannot be formally excluded, but does not seem to affect the results, as the frequency of PE⁺ memory cells remains constant from the time of their

Table 1 *in vitro* secondary antibody responses of PE⁺IgG1⁺ memory B cells

		Antibodies ($\mu\text{g ml}^{-1}$)					
		Anti-PE			Anti-NP		
Memory B cells	CD4 ⁺ T cells	IgG1	IgM	IgA	IgG1	IgM	IgA
PE ⁺ NP ⁻	PE primed	34.1 $\pm$ 1.0	<0.1	33.7 $\pm$ 0.1	<0.1	<0.1	<0.1
PE ⁺ NP ⁻	NP/CG primed	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1
NP ⁺	PE ⁻	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1
	primed						
		<0.1					

Numbers represent the concentration of anti-PE or anti-NP antibody (in $\mu\text{g ml}^{-1}$). Titres were obtained from three independent wells and the mean $\pm$ s.e.m. is shown.

appearance (Fig. 3c.) Given these circumstances, we believe that our data exclude the possibility that the maintenance of memory B cells *in vivo* depends on the occasional (high-affinity) interaction with persisting immunizing antigen, stored in the animal in an immunogenic form.

We conclude that the antibody system is able to generate true memory cells, which are independent of persisting immunizing antigen and enable the animal to produce secondary, high-affinity antibody responses. Like many naive B cells¹⁷, these cells are long-lived and act in concert with long-lived plasma cells secreting antibodies over extended periods of time^{20,21} and with memory T cells, for the maintenance of which antigen also does not seem to be required^{22–24}.

Methods

Anti-PE hybridomas

Splenic B cells from 4-month-old CkT homozygous mice²⁵ immunized intraperitoneally with 100 μg alum-precipitated PE were fused with X63.653 myeloma cells. Hybridomas secreting PE-specific IgG/1 antibodies with no detectable crossreactivity to NP₁₄-BSA were identified by enzyme-linked immunosorbent assay (ELISA).

Gene targeting and generation of *VNP/VPEinv*, *Mx-cre* mice

We cloned *VDP/PE* and its leader exon by PCR amplification using degenerate VH family primers and JH1-4 primers from hybridoma 80A3, and linked it to the VH promoter of the *VHBI-8* gene (PE cassette). The targeting vector was generated by inserting an 11-kilobase (kb) genomic DNA upstream of the DQ52 element with an *HSV-tk* gene, followed by a *tk-neo* (*neo*^r) cassette with a single *loxP*511 site, a reversed 3.3-kb PE cassette flanked by *loxP*511 sites in opposite orientation, and a 0.8-kb JH region into pBluescriptII (Stratagene). The targeting construct was electroporated into E14.1 embryonic stem (ES) cells. Procedures to obtain the appropriately targeted ES clones were essentially done as described²⁶. The targeted ES clones were aggregated with CD1-derived blastocysts, and the resulting chimaeric mice were bred for germline transmission to obtain *VPEinv* mice. *VNP/VPEinv*, *Mx-cre* mice were generated by intercrossing the *VNP/VNP*, *Mx-cre*¹⁰ and *VPEinv/VPEinv* mice. *VPE* mice were generated by crossing *VPEinv* mice to the deleter strain²⁷.

Cell purification and flow cytometric analysis

We collected spleens at various times after IFN injection, and prepared single-cell suspensions. For positive enrichment of the IgG1 population, splenocytes (0.7–1 $\times$ 10⁶ cells) were stained with a digoxigenin (DIG)-conjugated rat anti-mouse IgG1 monoclonal antibody²⁸, followed by anti-DIG conjugated MACS beads. Subsequently, the cells were first incubated with NP14-BSA (5 $\mu\text{g ml}^{-1}$) and biotinylated anti-mouse IgM (R33-24-12) followed by FITC-sheep anti-DIG Fab fragment (Boehringer), PE (10 $\mu\text{g ml}^{-1}$; Cyanotech), streptavidin-Texas Red or Alexa Fluor594 (Molecular Probes), and allophycocyanin (APC)-anti-NP monoclonal antibody S43 (ref. 28). Cells were resuspended in propidium iodide and analysed or sorted using a FACStar (Becton Dickinson).

Somatic mutation analysis of single cells

Single cells were sorted from pooled splenic cells of three experimental mice (Fig. 4). Genomic DNA amplification from single cells, and subsequent DNA sequencing were done as described²⁹. Briefly, to amplify genomic *VPEinv* DNA segments, we carried out two rounds of PCR with an annealing temperature of 55 °C. We used a 5' leader exon primer, V186.2L (5'-AGTTACTGAGCACACAGGACCT-3'), and a 3' JC intron primer, SAHIII (5'-TAATCTGTCTCTAAAGGCTCTGA-3'). The second round amplification was performed using 5' V186.2L primer and 3' JH3i (5'-CCCCAATGCATTGATGATGG-3'). For RT-PCR, complementary DNA was synthesized and amplified using TITAN kit (Boehringer). For analysis of *VPE* in PE⁺IgG1⁺ cells, we used a 3' Cy1 primer, IgG1CH1 (5'-ATGGAGTTA GTTTGGGAGCAGAT-3') for both cDNA synthesis and first-round amplification along with the 5' V186.2L primer. The second-round amplification was carried out using a set of nested 5' VHA²⁹ and 3' JH3PE (5'-GCAGAGACAGTGACCAG

AGT-3') primers. The reaction consisted of an initial denaturation at 94 °C for 2 min, followed by 30 cycles at 94 °C for 30 s each, 50 °C for 30 s and 68 °C for 45 s, with a final elongation step at 68 °C for 7 min.

In vitro secondary antibody response

PE⁺NP⁻ or NP⁺PE IgG1⁺ cells (250 cells per well) from pooled memory B cells from three experimental mice were sorted directly into 96-well plates containing splenic CD4⁺ T cells (2 $\times$ 10⁵ cells per well) from unimmunized, or from PE- or NP/CG-immunized (100 μg alum-precipitated antigen per immunization) C57BL/6 mice, purified by using anti-B220 and anti-CD4 (L3T4) MACS beads sequentially to more than 95% purity. Mixed T and B cells were cultured in RPMI1640 (Gibco) medium with 10% FCS for 14 days at 37 °C. Supernatants were analysed for the presence of PE or NP binding IgG1 antibodies by ELISA using PE or NP14-BSA coated (10 $\mu\text{g ml}^{-1}$) plates, followed by a biotin-goat anti-mouse IgG1, IgA antiserum or anti-mouse IgM monoclonal antibody as a secondary reagent. The concentration of anti-PE antibodies was determined by using either purified anti-PE 80A3 monoclonal antibody, or mouse anti-PE serum obtained day 14 after immunization as a standard. N1G9 and B1-8 μ monoclonal antibodies³⁰ were used as standards for anti-NP antibody titrations.

BrdU labelling and immunocytochemical staining

Memory B cells (PE⁺IgG1⁺; 2 $\times$ 10³ cells) and progenitor B cells (B220⁺IgM⁻; 5 $\times$ 10⁴ cells) were sorted from spleen and bone marrow of 3 BrdU-fed (1 mg ml⁻¹) experimental mice, respectively. In experiment 1, total IgG1⁺ memory B cells (B220⁺IgG1⁺; 1 $\times$ 10⁴ cells) were also sorted. The sorted PE⁺ memory B cells were mixed with Ficoll-isolated J558 murine myeloma cells (4 $\times$ 10⁴ cells) as carriers and attached to slides using Cytospin. After air-drying for 5 min, the cells fixed on the slides for 30 min at -20 °C with 5% glacial acetic acid in ethanol were treated with 1.5 M HCl for 20 min for denaturation of DNA, and then neutralized twice in 0.1 M disodium tetraborate, pH 9.0. Treated cells were incubated with biotinylated anti-BrdU antibody, PRB-1 (12.5 $\mu\text{g ml}^{-1}$, Alexis) for 30 min, washed in PBS, and labelled cells were determined with streptavidin-Alexa Fluor488 (Molecular Probes) after incubation for 20 min. Stained slides were mounted in Fluoromount G (Southern Biotechnology Associates) and counted under a fluorescence microscope.

Received 8 May; accepted 17 July 2000.

- Sprent, J. T and B memory cells. *Cell* **76**, 315–322 (1994).
- Ahmed, R. & Gray, D. Immunological memory and protective immunity: understanding their relation. *Science* **272**, 54–60 (1996).
- Mandel, T. E., Phipps, R. P., Abbot, A. & Tew, J. G. The follicular dendritic cell: long term antigen retention during immunity. *Immunol. Rev.* **53**, 29–59 (1980).
- Askonas, B. A., Cunningham, A. J., Kreth, H. W., Roelants, G. E. & Williamson, A. R. Amplification of B cell clones forming antibody to the 2,4-dinitrophenyl group. *Eur. J. Immunol.* **2**, 494–498 (1972).
- Gray, D. & Skarvall, H. B-cell memory is short-lived in the absence of antigen. *Nature* **336**, 70–73 (1988).
- Schitteck, B. & Rajewsky, K. Maintenance of B-cell memory by long-lived cells generated from proliferating precursors. *Nature* **346**, 749–751 (1990).
- Vieira, P. & Rajewsky, K. Persistence of memory B cells in mice deprived of T cell help. *Int. Immunol.* **2**, 487–494 (1990).
- Karrer, U. et al. Antiviral B cell memory in the absence of mature follicular dendritic cell networks and classical germinal centers in TNFR1^{-/-} mice. *J. Immunol.* **164**, 768–778 (2000).
- Lam, K. P. & Rajewsky, K. Rapid elimination of mature autoreactive B cells demonstrated by Cre-induced change in B cell antigen receptor specificity *in vivo*. *Proc. Natl Acad. Sci. USA* **95**, 13171–13175 (1998).
- Lam, K. P., Kühn, R. & Rajewsky, K. *In vivo* ablation of surface immunoglobulin on mature B cells by inducible gene targeting results in rapid cell death. *Cell* **90**, 1073–1083 (1997).
- Hayakawa, K., Ishii, R., Yamasaki, K., Kishimoto, T. & Hardy, R. R. Isolation of high-affinity memory B cells: phycoerythrin as a probe for antigen-binding cells. *Proc. Natl Acad. Sci. USA* **84**, 1379–1383 (1987).
- Hoess, R. H., Wierzbicki, A. & Abremski, K. The role of the *loxP* spacer region in P1 site-specific recombination. *Nucleic Acids Res.* **14**, 2287–2300 (1986).
- Kühn, R., Schwenk, F., Aguet, M. & Rajewsky, K. Inducible gene targeting in mice. *Science* **269**, 1427–1429 (1995).
- Okumura, K., Julius, M. H., Tsu, T., Herzenberg, L. A. & Herzenberg, L. A. Demonstration that IgG memory is carried by IgG-bearing cells. *Eur. J. Immunol.* **6**, 467–472 (1976).
- Esser, C. & Radbruch, A. Immunoglobulin class switching: molecular and cellular analysis. *Annu. Rev. Immunol.* **8**, 717–735 (1990).
- MacLennan, I. C. M. Germinal centers. *Annu. Rev. Immunol.* **12**, 117–139 (1994).
- Rajewsky, K. Clonal selection and learning in the antibody system. *Nature* **381**, 751–758 (1996).
- Neuberger, M. S. et al. Monitoring and interpreting the intrinsic features of somatic hypermutation. *Immunol. Rev.* **162**, 107–116 (1998).
- Nemazee, D. Antigen receptor 'capacity' and the sensitivity of self-tolerance. *Immunol. Today* **17**, 25–29 (1996).
- Manz, R. A., Thiel, A. & Radbruch, A. Lifetime of plasma cells in the bone marrow. *Nature* **388**, 133–134 (1997).
- McHeyzer-Williams, M. G. & Ahmed, R. B cell memory and the long-lived plasma cell. *Curr. Opin. Immunol.* **11**, 172–179 (1999).
- Bruno, L., Kirberg, J. & von Boehmer, H. On the cellular basis of immunological T cell memory. *Immunity* **2**, 37–43 (1995).
- Murali-Krishna, K. et al. Persistence of memory CD8 T cells in MHC class I-deficient mice. *Science* **286**, 1377–1381 (1999).
- Swain, S. L., Hu, H. & Huston, G. Class II-independent generation of CD4 memory T cells from effectors. *Science* **286**, 1381–1383 (1999).

25. Zou, Y. R., Takeda, S. & Rajewsky, K. Gene targeting in the Ig κ locus: efficient generation of λ chain-expressing B cells, independent of gene rearrangements in Ig κ . *EMBO J.* **12**, 811–820 (1993).
26. Torres, R. M. & Kühn, R. *Laboratory Protocols for Conditional Gene Targeting* (Oxford Univ. Press, Oxford, 1997).
27. Schwenk, F., Baron, U. & Rajewsky, K. A *cre*-transgenic mouse strain for the ubiquitous deletion of *loxP*-flanked gene segments including deletion in germ cells. *Nucleic Acids Res.* **23**, 5080–5081 (1995).
28. Kaisho, T., Schwenk, F. & Rajewsky, K. The roles of $\gamma 1$ heavy chain membrane expression and cytoplasmic tail in IgG1 responses. *Science* **276**, 412–415 (1997).
29. Ehlich, A., Martin, V., Müller, W. & Rajewsky, K. Analysis of the B-cell progenitor compartment at the level of single cells. *Curr. Biol.* **4**, 573–583 (1994).
30. Cumano, A. & Rajewsky, K. Structure of primary anti-(4-hydroxy-3-nitrophenyl)acetyl (NP) antibodies in normal and idiotypically suppressed C57BL/6 mice. *Eur. J. Immunol.* **15**, 512–520 (1985).

Acknowledgements

We are grateful to B. Hampel, A. Roth and C. Göttlinger for expert technical help, and to T. Kaisho, S. Schwes, L. Pao, A. Ehlich, and R. Küppers for discussion. We also thank S. Casola, P. Oberdoerffer and W. Müller for comments on the manuscript and G. Schmall for help in its preparation. This work was supported by the Volkswagen Foundation, the European Union and the Deutsche Forschungsgemeinschaft. M.M. was recipient of a fellowship from the Alexander von Humboldt Foundation, and K.-P.L. was a fellow of the Human Frontier Science Organization.

Correspondence and requests for materials should be addressed to K.R. (e-mail: klaus.rajewsky@genetik.uni-koeln.de). The GenBank accession number for the *VDJPE* gene sequence is AF262991.

A common E2F-1 and p73 pathway mediates cell death induced by TCR activation

Natalie A. Lissy*, Penny K. Davis*, Meredith Irwin†, William G. Kaelin† & Steven F. Dowdy*

* Howard Hughes Medical Institute, Departments of Pathology and Medicine, Washington University School of Medicine, St. Louis, Missouri 63110, USA;

† Howard Hughes Medical Institute, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts 02115, USA

Strong stimulation of the T-cell receptor (TCR) on cycling peripheral T cells causes their apoptosis by a process called TCR-activation-induced cell death (TCR-AICD)^{1–3}. TCR-AICD occurs from a late G1 phase cell-cycle check point⁴ independently of the ‘tumour suppressor’ protein p53 (refs 5, 6). Disruption of the gene for the E2F-1 transcription factor^{7,8}, an inducer of apoptosis^{9–11}, causes significant increases in T-cell number and splenomegaly^{12–15}. Here we show that T cells undergoing TCR-AICD induce the p53-related gene p73, another mediator of apoptosis¹⁶, which is hypermethylated in lymphomas^{17,18}. Introducing a dominant-negative E2F-1 protein or a dominant-negative p73 protein into T cells protects them from TCR-mediated apoptosis, whereas dominant-negative E2F-2, E2F-4 or p53 does not. Furthermore, E2F-1-null or p73-null primary T cells do not undergo TCR-mediated apoptosis either. We conclude that TCR-AICD occurs from a late G1 cell-cycle checkpoint that is dependent on both E2F-1 and p73 activities. These observations indicate that, unlike p53, p73 serves to integrate receptor-mediated apoptotic stimuli.

Previously, we showed that TCR-AICD occurs from a late G1 cell-cycle death checkpoint⁴ in a p53-independent manner^{5,6}. However, the mediators of this apoptotic programme remain unknown. To examine this process, we used centrifugally elutriated early G1 phase Jurkat T cells and mimicked TCR engagement by crosslinking the TCR with anti-CD3 antibodies¹⁹. Apoptosis was first detected in

TCR-stimulated early G1 T cells at 6 h after stimulation by trypan blue vital dye exclusion and TdT-mediated dUTP nick end labelling (TUNEL) assay (Fig. 1a). In addition, treatment with cycloheximide, a protein synthesis inhibitor, inhibited TCR-mediated apoptosis (data not shown), indicating that translation and probably transcription were required.

Mice without the gene for the transcription factor E2F-1 suffer from splenomegaly enlarged lymph nodes and lymphomas, potentially owing to a defect in apoptosis^{12,13}. Also, overexpressed ectopic E2F-1 induces apoptosis^{9–11}. In addition, the late G1 cell-cycle position of TCR-mediated apoptosis coincides with transcriptional activation by E2F family members^{7,8}. To examine the involvement of E2F-1 as a mediator of TCR-AICD, we generated transducible TAT-fusion proteins of a dominant-negative E2F-1 protein (residues 1–374) lacking the carboxy-terminal transcriptional transactivation domain (TAT–E2F-1^{DN}), an inactive DNA-binding domain (DBD) mutant (TAT–E2F-1^{ΔDBD}), and a TAT–green fluorescent protein (GFP) fusion as a negative control (Fig. 1b, top). TAT–E2F-1^{DN} protein bound to DNA containing an E2F site *in vitro*, but the TAT–E2F-1^{ΔDBD} protein did not (data not shown). TAT-fusion proteins transduce into about 100% of cells in a rapid, concentration-dependent manner, reaching intracellular equilibrium in less than 15 min^{4,20}, and both TAT–E2F-1^{DN} and TAT–E2F-1^{ΔDBD} fusion proteins were found to transduce into T cells (Fig. 1b, bottom).

TCR-stimulated early G1 phase T cells were treated with 100 nM TAT–E2F-1^{DN}, TAT–E2F-1^{ΔDBD} or TAT–GFP proteins and assayed for apoptosis by TUNEL assay (Fig. 1c) and trypan blue dye exclusion. T cells treated with PBS, TAT–GFP and TAT–E2F-1^{ΔDBD} proteins showed similar high levels of TUNEL-positive genomic DNA at 6 and 9 h after stimulation. In contrast, T cells treated with TAT–E2F-1^{DN} protein were resistant to both DNA fragmentation and apoptosis (Fig. 1c). Furthermore, stimulated T cells treated

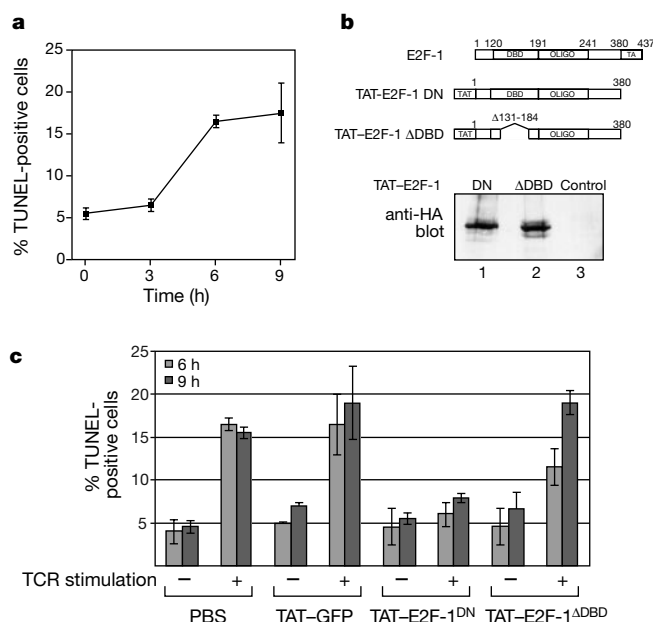


Figure 1 TCR-AICD occurs from late G₁ phase and is E2F-1-dependent. **a**, Early G₁ elutriated human Jurkat T cells were TCR (anti-CD3) stimulated and assayed for apoptosis by TUNEL genomic DNA fragmentation. **b**, Diagram of transducible TAT–E2F-1 dominant-negative (TAT–E2F-1^{DN}) and E2F-1 DNA-binding domain (TAT–E2F-1^{ΔDBD}) mutants (top). Anti-HA immunoblot analysis from T cells treated with 100 nM TAT–E2F-1 fusion proteins (bottom). Note that the HA epitope is present in the TAT leader. **c**, Parallel cultures of TCR (anti-CD3)-stimulated (+) or unstimulated (–) early G₁ T cells were treated with 100 nM TAT–E2F-1^{DN}, TAT–E2F-1^{ΔDBD} or control TAT–GFP proteins and assayed for apoptosis at 3, 6 and 9 h after treatment by TUNEL genomic DNA fragmentation. All experiments were performed in duplicate and error bars indicate s.e.m.

with dominant-negative TAT-E2F-2^{DN} and TAT-E2F-4^{DN} proteins were not protected from apoptosis (data not shown). These observations indicate that TCR-AICD is an E2F-1-dependent process.

TCR-AICD is independent of p53 (refs 5, 6, 21), but a p53 family member, p73, induces apoptosis when overexpressed¹⁶. Ectopic

expression of E2F-1 induces p73 messenger RNA levels and apoptosis²². We therefore examined the amount of p73 in TCR-stimulated early G1 T cells over a time course by immunoblot analysis (Fig. 2). p73 levels remained relatively unchanged in control T cells. In contrast, p73 amounts began to increase by 3 h after TCR

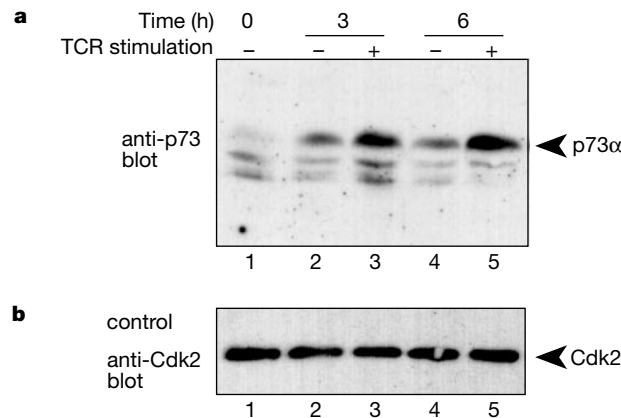


Figure 2 TCR-AICD induces p73 protein levels. **a**, Anti-p73 immunoblot analysis from parallel cultures of TCR (anti-CD3) stimulated (+) or unstimulated (-) early G₁ human T cells. p73α form (relative molecular mass 80,000) is the predominant form observed in

T cells. **b**, Control anti-Cdk2 immunoblot to normalize protein loading. Unstimulated and stimulated T cells contain constant levels of Cdk2 (ref. 4).

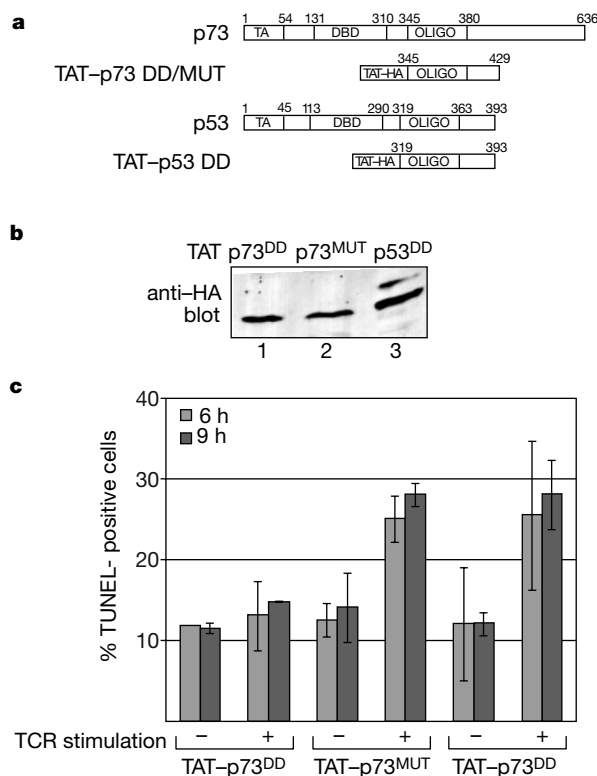


Figure 3 TCR-AICD is p73-dependent. **a**, Diagram of transducible TAT-p73 dominant-negative (TAT-p73^{DD}), TAT-p73 mutant (TAT-p73^{MUT}) and TAT-p53 dominant-negative (TAT-p53^{DD}) proteins. **b**, Anti-HA immunoblot analysis from T cells treated with 100 nM TAT-p73 and TAT-p53 fusion proteins. Asterisk, L371P mutation in TAT-p73^{MUT}. **c**, TCR (anti-CD3) stimulated elutriated early G₁ human T cells were treated with 100 nM TAT-p73^{DD}, TAT-p73^{MUT} or control TAT-p53^{DD} proteins and assayed for apoptosis by TUNEL genomic DNA fragmentation at 6 and 9 h after addition. All experiments were performed in duplicate and error bars indicate s.e.m.

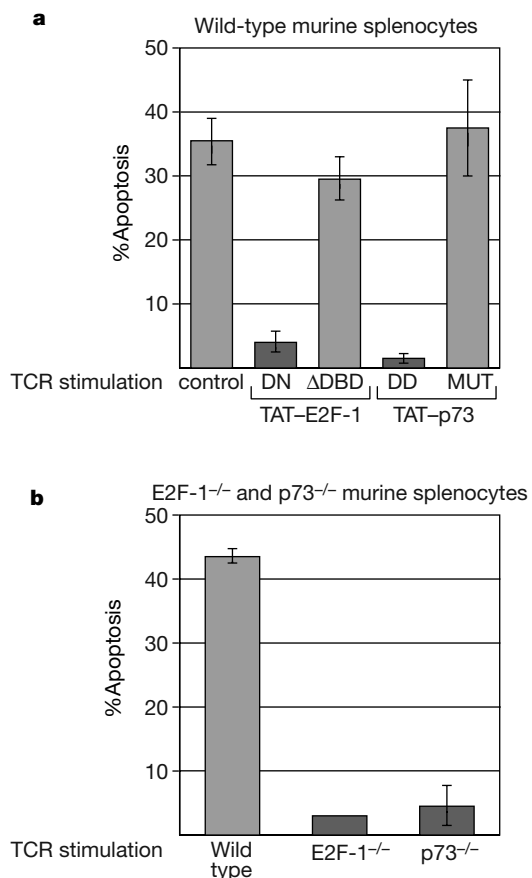


Figure 4 Primary T cells require E2F-1 and p73 to undergo TCR-AICD. **a**, TCR (anti-CD3)-stimulated asynchronous, cycling primary wild-type murine peripheral T cells (splenocytes) were treated 100 nM TAT-E2F-1^{DN}, control TAT-E2F-1^{ΔDBD}, TAT-p73^{DD} or control TAT-p73^{MUT} proteins and assayed for apoptosis at 20 h after addition. All experiments were performed in triplicate and error bars indicate s.e.m. **b**, Cycling wild-type, E2F-1-null and p73-null primary T cells (splenocytes) were TCR (anti-CD3) stimulated and assayed for apoptosis at 20 h after stimulation. Error bars indicate s.e.m.

stimulation, slightly before the increase of TUNEL-positive T cells, and are consistent with a recent report²³. To investigate the role of p73 in mediating TCR-AICD, we generated a transducible p73 dominant-negative TAT-fusion protein (TAT-p73^{DD}) corresponding to the tetramerization domain of p73 (residues 329–430) that binds to and interferes with p73 function, but does not bind p53 (refs 22, 24; Fig. 3a). An inactive control TAT-p73^{MUT} mutant protein, corresponding to an inactivating tetramerization point mutation (L371P) found in p53 from Li-Fraumeni patients^{25,26}, and a p53 tetramerization domain dominant-negative (TAT-p53^{DD}) control protein were also generated (Fig. 3a). TAT-p73^{DD}, TAT-p73^{MUT} and TAT-p53^{DD} proteins transduced into cells with near equal efficiency (Fig. 3b). TCR-stimulated early G1 T cells were treated with 100 nM TAT-p73^{DD}, TAT-p73^{MUT} or TAT-p53^{DD} proteins and assayed for apoptosis by TUNEL genomic DNA fragmentation at 6 and 9 h after addition (Fig. 3c). Both TAT-p73^{MUT} and TAT-p53^{DD} proteins failed to protect T cells from apoptosis. In contrast, TCR-stimulated T cells treated with TAT-p73^{DD} protein showed only background levels of TUNEL-positive cells.

To determine further the requirement for E2F-1 and p73 in T-cell apoptosis, we examined anti-CD3-stimulated primary murine peripheral T cells (splenocytes). Consistent with the observations above, stimulated murine splenocytes were also protected from apoptosis by treatment with TAT-E2F-1^{DN} and TAT-p73^{DD} proteins, whereas inactive mutant proteins failed to inhibit apoptosis (Fig. 4a). We next assayed the ability of E2F-1-null^{12,13} and p73-null²⁷ primary splenocytes to undergo apoptosis. Anti-CD3-stimulated wild-type T cells underwent apoptosis consistent with the results above (Fig. 4b). In contrast, both the E2F-1-null and the p73-null cells failed to undergo apoptosis (Fig. 4b). These observations from dominant-negative transducing proteins and genetic deletions show that E2F-1 and p73 are specific mediators of TCR-AICD.

Previous studies have demonstrated that ectopic overexpression of E2F-1 promotes apoptosis^{9–11}, but here we show that TCR-mediated apoptosis requires endogenous levels of E2F-1 to initiate apoptosis in mature peripheral T cells. These observations are consistent with the disruption of lymphatic homeostasis observed in E2F-1-null mice^{12–15}. TCR-AICD is a p53-independent process^{5,6,21}, but the p53-related gene, p73, is both induced in T cells undergoing TCR-AICD²³ and a downstream target of E2F-1 (ref. 22). Consistent with these observations, inactivation of pRB, which is a regulator of E2F-1, by SV40 T antigen or adenoviral E1A also results in increased p73 (refs 28, 29). Moreover, several studies have reported transcriptional silencing of the p73 gene by hypermethylation in human lymphomas, indicating a possible involvement in oncogenesis^{17,18}. Thus, based on the ability of TAT-p73^{DD} protein to block apoptosis and the absence of apoptosis in both E2F-1- and p73-null T cells, we conclude that TCR-mediated apoptosis is dependent on both E2F-1 and p73. These observations indicate that a common pathway mediates TCR-AICD where p73 is a downstream target of E2F-1, rather than E2F-1 induction of an unidentified target gene. However, we cannot exclude the possibility that E2F-1 and p73 have parallel, but obligate pathways to induce apoptosis. Thus, unlike p53, which serves as a checkpoint protein for DNA damage, p73 appears to integrate stimuli for both DNA damage and receptor-mediated apoptosis. □

Methods

Cell culture

Human Jurkat T cells were obtained from ATCC and maintained as previously described⁴. Centrifugal elutriations were performed as described⁴. Murine splenocytes were isolated as described³⁰ and activated by treatment with con-A (2 µg ml⁻¹) and IL-2 (10 U ml⁻¹). For TCR stimulation, elutriated T cells were resuspended in fresh media at 1.5 × 10⁶ cells ml⁻¹ and treated with 4 µg ml⁻¹ of soluble anti-CD3 (HIT3a) antibodies (Pharmingen) and splenocytes were treated with plate crosslinked hamster anti-murine CD3 antibodies (gift from J. Russell) plus IL-2 (10 U ml⁻¹). TAT-fusion proteins were added to T-cell cultures

one hour before stimulation at 100 nM final concentration. Cell viability was determined by TUNEL positive genomic DNA assay (Trevigen) as described by the manufacturer and trypan blue exclusionary dye. The percentage of TUNEL-positive cells was determined by dividing the number of TUNEL-positive staining cells per field by total number of cells. Confocal microscopy of cells treated with fluorescein-isothiocyanate-labelled TAT-fusion proteins was performed as described²⁰.

TAT-fusion protein purification

E2F-1 complementary DNAs (residues 1–374 and 1–374Δ131–184) were cloned into the NcoI–EcoRI site of pTAT-HA vector²⁰. p73 dominant-negative and inactive point mutant (residues 326–430), and p53 dominant-negative (residues 319–393) cDNAs were cloned into the KpnI–EcoRI sites of the pTAT-HA vector. TAT–GFP was generated by inserting the GFP open reading frame into the NcoI and EcoRI sites of pTAT-HA. All TAT-fusion proteins were purified as described²⁰, fractions containing TAT-fusion protein underwent buffer exchange on PD10 columns (Pharmingen) into PBS (TAT–E2F-1) or TE (TAT–p73, p53). Immunoblots were performed with anti-haemagglutinin (anti-HA; Babco), anti-p73 (Oncogene) or anti-Cdk2 (Santa Cruz) antibodies as described²⁰.

Received 7 July; accepted 2 August 2000.

- Jones, L. A., Chin, L. T., Longo, D. L. & Kruisbeek, A. M. Peripheral clonal elimination of functional T cells. *Science* **250**, 1726–1729 (1990).
- Rocha, B. & von Boehmer, H. Peripheral selection of the T cell repertoire. *Science* **251**, 1225–1228 (1991).
- Kabelitz, D., Pohl, T. & Pechhold, K. Activation-induced cell death (apoptosis) of mature peripheral T lymphocytes. *Immunol. Today* **14**, 338–339 (1993).
- Lissy, N. A. et al. TCR antigen-induced cell death occurs from a late G1 phase cell cycle check point. *Immunity* **8**, 57–65 (1998).
- Boehme, S. A. & Lenardo, M. J. TCR-mediated death of mature T lymphocytes occurs in the absence of p53. *J. Immunol.* **156**, 4075–4078 (1996).
- Malcomson, R. D. et al. Apoptosis induced by gamma-irradiation, but not CD4 ligation, of peripheral T lymphocytes in vivo is p53-dependent. *J. Pathol.* **181**, 166–171 (1997).
- Dyson, N. The regulation of E2F by pRB-family proteins. *Genes Dev.* **12**, 2245–2262 (1998).
- Nevins, J. R. Toward an understanding of the functional complexity of the E2F and retinoblastoma families. *Cell Growth Differ.* **9**, 585–593 (1998).
- Qin, X. Q., Livingston, D. M., Kaelin, W. G. & Adams, P. D. Deregulated transcription factor E2F-1 expression leads to S-phase entry and p53-mediated apoptosis. *Proc. Natl Acad. Sci. USA* **91**, 10918–10922 (1994).
- Wu, X. & Levine, A. J. p53 and E2F-1 cooperate to mediate apoptosis. *Proc. Natl Acad. Sci. USA* **91**, 3602–3606 (1994).
- DeGregori, J., Leone, G., Miron, A., Jakoi, L. & Nevins, J. R. Distinct roles for E2F proteins in cell growth control and apoptosis. *Proc. Natl Acad. Sci. USA* **94**, 7245–7250 (1997).
- Yamasaki, L. et al. Tumor induction and tissue atrophy in mice lacking E2F-1. *Cell* **85**, 537–548 (1996).
- Field, S. J. et al. E2F-1 functions in mice to promote apoptosis and suppress proliferation. *Cell* **85**, 549–561 (1996).
- Zhu, J. W., DeRyckere, D., Li, F. X., Wan, Y. Y. & DeGregori, J. A role for E2F-1 in the induction of ARF, p53, and apoptosis during thymic negative selection. *Cell Growth Differ.* **10**, 829–838 (1999).
- Garcia, I., Murga, M., Vicario, A., Field, S. J. & Zubiaga, A. M. A role for E2F-1 in the induction of apoptosis during thymic negative selection. *Cell Growth Differ.* **11**, 91–98 (2000).
- Jost, C. A., Marin, M. C. & Kaelin, W. G. p73 is a human p53-related protein that can induce apoptosis. *Nature* **389**, 191–194 (1997).
- Corn, P. G. Transcriptional silencing of the p73 gene in acute lymphoblastic leukemia and Burkitt's lymphoma is associated with 5' CpG island methylation. *Cancer Res.* **59**, 3352–3356 (1999).
- Kawano, S. et al. Loss of p73 gene expression in leukemias/lymphomas due to hypermethylation. *Blood* **94**, 1113–1120 (1999).
- Smith, C. A., Williams, G. T., Kingston, R., Jenkinson, E. J. & Owen, J. J. Antibodies to CD3/T-cell receptor complex induce death by apoptosis in immature T cells in thymic cultures. *Nature* **337**, 181–184 (1989).
- Nagahara, H. et al. Transduction of full-length TAT fusion proteins into mammalian cells: TAT-p27Kip1 induces cell migration. *Nature Med.* **4**, 1449–1452 (1998).
- Donehower, L. A. et al. Mice deficient for p53 are developmentally normal but susceptible to spontaneous tumours. *Nature* **356**, 215–221 (1992).
- Irwin, M. et al. Role for the p53 homologue p73 in E2F-1-induced apoptosis. *Nature* **407**, 645–648 (2000).
- De Laurenzi, V. D. et al. Additional complexity in p73: induction by mitogens in lymphoid cells and identification of two new splicing variants epsilon and zeta. *Cell Death Differ.* **6**, 389–390 (1999).
- Kaelin, W. G. The p53 gene family. *Oncogene* **18**, 7701–7705 (1999).
- Lomax, M. E., Barnes, D. M., Hupp, T. R., Pickles, S. M. & Camplejohn, R. S. Characterization of p53 oligomerization domain mutations isolated from Li-Fraumeni and Li-Fraumeni like family members. *Oncogene* **17**, 643–649 (1998).
- Davison, T. S., Yin, P., Nie, E., Kay, C. & Arrowsmith, C. H. Characterization of the oligomerization defects of two p53 mutants found in families with Li-Fraumeni and Li-Fraumeni-like syndrome. *Oncogene* **17**, 651–656 (1998).
- Yang, A. p73-deficient mice have neurological, pheromonal and inflammatory defects but lack spontaneous tumours. *Nature* **404**, 99–103 (2000).
- Kovalev, S., Marchenko, N., Swendeman, S., LaQuaglia, M. & Moll, U. M. Expression level, allelic origin, and mutation analysis of the p73 gene in neuroblastoma tumors and cell lines. *Cell Growth Differ.* **9**, 897–903 (1998).
- Marin, M. C. et al. Viral oncoproteins discriminate between p53 and the p53 homolog p73. *Mol. Cell. Biol.* **18**, 6316–6324 (1998).
- Abrams, S. I. & Russell, J. H. CD4+ T lymphocyte-induced target cell detachment. A model for T cell-mediated lytic and nonlytic inflammatory processes. *J. Immunol.* **146**, 405–413 (1991).

Acknowledgements

We thank J. Lees and F. McKeon for providing E2F-1- and p73-null spleens, respectively, and J. Russell for anti-murine CD3 antibodies. This work was supported by the Howard Hughes Medical Institute.

Correspondence and requests for materials should be addressed to S.F.D. (e-mail: dowdy@pathology.wustl.edu).

Role for the p53 homologue p73 in E2F-1-induced apoptosis

Meredith Irwin*, Maria Carmen Marin*, Andrew C. Phillips†, Ratnam S. Seelan‡, David I. Smith‡, Wanguo Liu‡, Elsa R. Flores§, Kenneth Y. Tsai§, Tyler Jacks§, Karen H. Vousden† & William G. Kaelin Jr*||

* Dana-Farber Cancer Institute and Brigham and Womens Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA

† Regulation of Cell Growth Laboratory, NCI-FCRDC, Frederick, Maryland 21702, USA

‡ Department of Laboratory Medicine and Pathology, Mayo Clinic/ Mayo Medical School, Rochester, Minnesota 55905, USA

§ Department of Biology and Center for Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

|| Howard Hughes Medical Institute, Chevy Chase, Maryland 20185-6789, USA

The transcription factor E2F-1 induces both cell-cycle progression and, in certain settings, apoptosis. E2F-1 uses both p53-dependent and p53-independent pathways to kill cells^{1–8}. The p53-dependent pathway involves the induction by E2F-1 of the human tumour-suppressor protein p14ARF, which neutralizes HDM2 (human homologue of MDM2) and thereby stabilizes the p53 protein⁹. Here we show that E2F-1 induces the transcription of the p53 homologue p73. Disruption of p73 function inhibited E2F-1-induced apoptosis in p53-defective tumour cells and in p53^{−/−} mouse embryo fibroblasts. We conclude that activation of p73 provides a means for E2F-1 to induce death in the absence of p53.

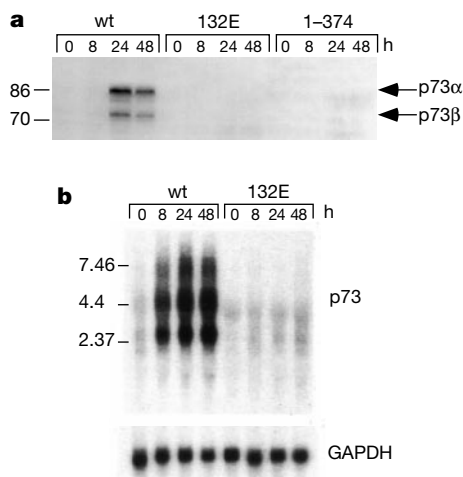


Figure 1 E2F-1 induces p73 mRNA and protein accumulation. **a**, SAOS2 cells were immunoblotted with an anti-p73 antibody at the indicated times after induction of wild-type (wt) or mutant E2F-1. **b**, Northern blot analysis of SAOS2 cells with radiolabelled p73 and GAPDH cDNA probes at the indicated time points after induction of E2F-1 or E2F-1(132E). Several p73 mRNA transcripts were detected, possibly owing to alternative splicing¹¹.

To determine whether E2F-1 could activate p73, we used SAOS2 osteosarcoma subclones in which wild-type or mutant derivatives of E2F-1 were controlled by an inducible promoter¹⁰. Induction of wild-type E2F-1 led to an increase in p73 protein levels, whereas induction of DNA-binding (E2F-1(132E)) or transactivation (E2F-1(1–374)) defective E2F-1 mutants did not (Fig. 1a). Similarly, p73 levels were increased after transient transfection of SAOS2 cells and C33A cervical carcinoma cells with plasmids encoding wild-type, but not mutant, E2F-1 (data not shown). Note that p73 and the related gene p63 both give rise to several protein isoforms owing to differential splicing¹¹. Both p73α and p73β can induce apoptosis¹¹.

E2F-1 activates the transcription of p14ARF, which neutralizes HDM2 and stabilizes p53 (ref. 9). In contrast, direct induction of p14ARF had no effect on p73 protein levels (data not shown). Furthermore, HDM2 degrades p53 but not p73 (ref. 11). Together, these findings suggested that induction of p73 by E2F-1 was not due to post-transcriptional changes mediated by ARF and HDM2. Indeed, E2F-1 induced the accumulation of p73 messenger RNA (Fig. 1b) and transcriptionally activated the p73 promoter (Fig. 2a). In parallel experiments, E2F-1 did not affect p63 mRNA or protein levels (data not shown) despite the high degree of similarity between p63 and p73 (ref. 11). We found that E2F-1 was the most potent activator of both the p73 promoter (Fig. 2b) and apoptosis (data not shown) among the E2Fs tested. In many systems, E2F-1 is the most potent inducer of apoptosis among the E2F family members¹². Inspection of the p73 promoter revealed at least three potential E2F-binding sites located near the transcription start site (see Supplementary Information). The authenticity of these sites was confirmed by electrophoretic mobility shift assays (data not shown), and their removal by a 5' deletion rendered the p73 promoter unresponsive to E2F-1 (see Supplementary Information). Inactivation of these three E2F sites by site-directed mutagenesis did not fully inactivate the promoter (data not shown), suggesting that the resulting mutant p73 promoter contains at least one non-canonical E2F-binding site or that E2F also has indirect effects on p73 transcription. Detailed characterization of the p73 promoter will be reported elsewhere.

E2F transcriptional activity is modulated by pRB family members and peaks in late G1 or S phase¹². We next measured p73 levels in cells that were synchronized by protocols that affect E2F activity. First, confluent HaCaT keratinocytes were plated at low density in the presence or absence of transforming growth factor (TGF)-β. TGF-β prevents pRB phosphorylation and thus suppresses E2F activity and cell-cycle progression. Likewise, TGF-β inhibited p73

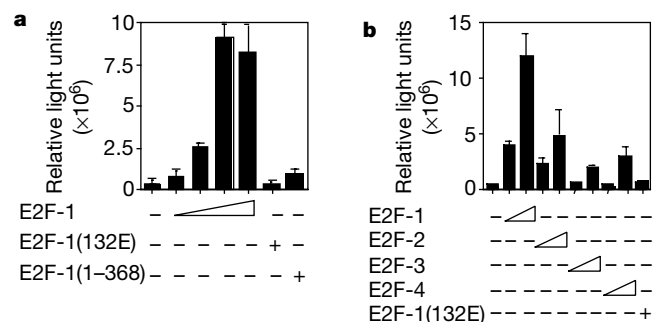


Figure 2 The p73 promoter is E2F-responsive. **a**, **b**, SAOS2 cells were transfected to produce the indicated haemagglutinin (HA)-tagged E2F proteins, along with a luciferase reporter plasmid containing the p73 promoter. Plasmid concentrations: triangle represents 10 ng, 100 ng, 1 μg and 2 μg in **a**, and 100 ng and 1 μg in **b**; '+' represents 2 μg in **a** and 1 μg in **b**. Each E2F protein was produced at comparable levels as determined by anti-HA immunoblot analysis (data not shown). Shown are luciferase values corrected for transfection efficiency ($n = 3$ independent experiments).

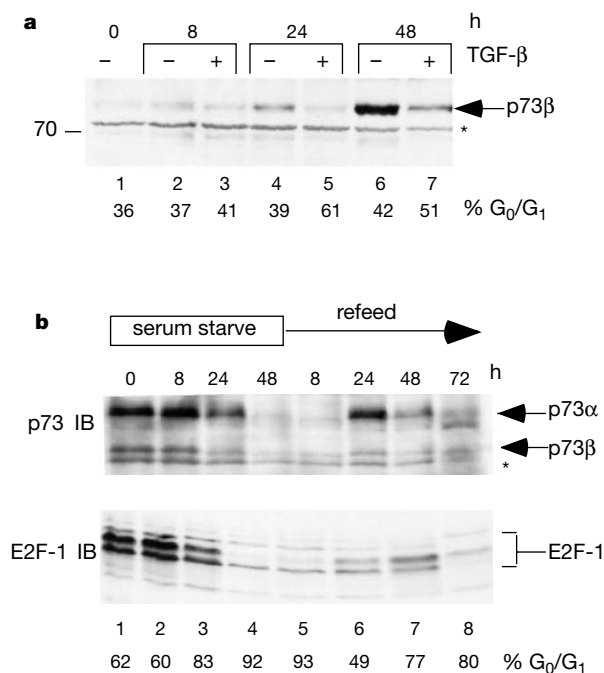


Figure 3 Modulation of p73 levels by physiological perturbants of E2F. **a**, Anti-p73 immunoblot analysis of HaCaT cells grown in the presence or absence of TGF- β for the indicated periods of time. **b**, Anti-p73 and anti-E2F-1 immunoblot (IB) analysis of T98G cells that were serum starved for 48 h and then re-fed. Each lane contained the same

amount of protein, as determined by Bradford assay. Nonspecific bands are indicated with an asterisk. The percentage of cells in G₀/G₁ was determined by propidium iodide staining followed by FACS.

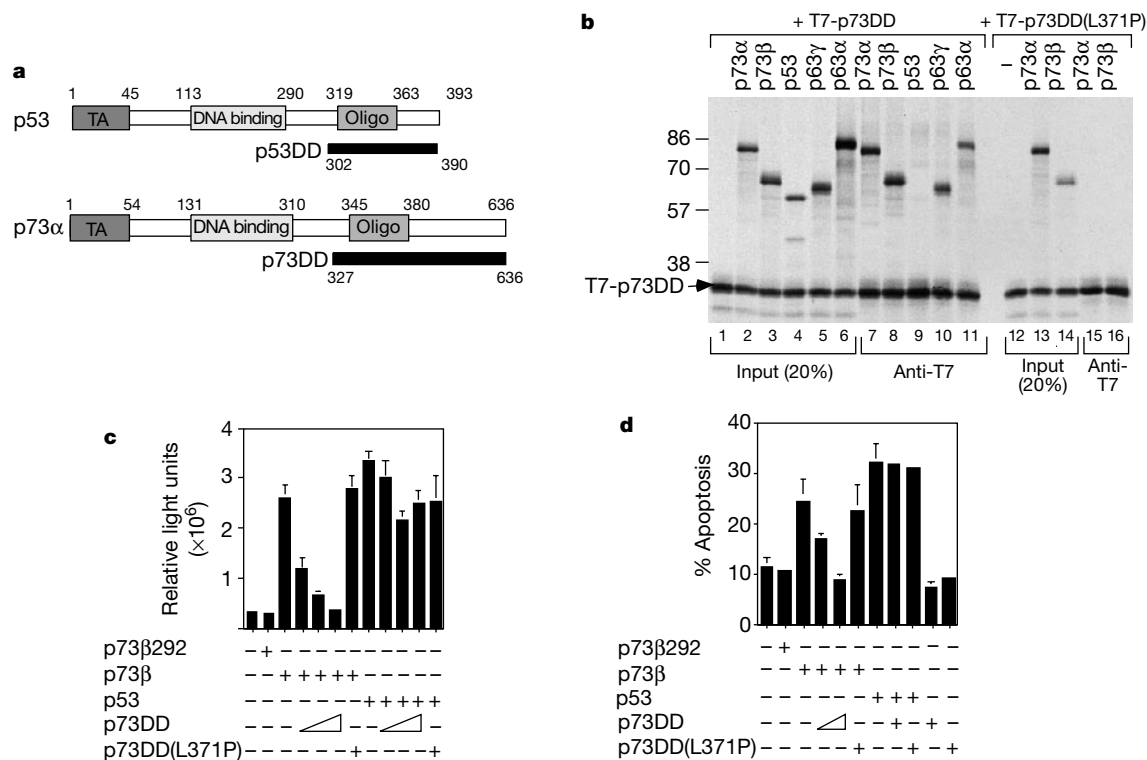


Figure 4 Validation of a dominant-negative p73 mutant. **a**, Representation of dominant-negative p73 mutant. **b**, T7-epitope tagged p73DD or T7-p73DD(L371P) was co-translated with the indicated p73, p53 and p63 isoforms in the presence of ³⁵S-methionine and immunoprecipitated with anti-T7 antibody. **c**, SAOS2 cells were transfected with a luciferase reporter plasmid containing the p21 promoter along with plasmids encoding wild-type p73, wild-type p53 or mutant p73 (R292H) (1 μ g) in the

presence or absence of increasing amounts of plasmids encoding p73DD (0.5, 1 or 2 μ g) or p73DD(L371P) (2 μ g). Luciferase values were normalized for transfection efficiency. **d**, SAOS2 cells were transfected with a plasmid encoding p73β or p53 (5 μ g) in the presence or absence of increasing amounts of a plasmids encoding p73DD (5 or 20 μ g) or p73DD(L371P) (20 μ g). The number of apoptotic cells ($n = 3$ independent experiments) was determined by FACS 48 h later. Error bar indicates 1 standard error.

accumulation (Fig. 3a). We next grew T98G glioblastoma cells in low serum (Fig. 3b). The levels of both E2F-1 and p73 fell as these cells accumulated in G0/1. Note that E2F-1 is itself transcribed by an E2F-responsive promoter¹². Again, p73 was detectable as cells exited G1 and entered S phase after the re-addition of serum. These results support the idea that p73 is regulated by E2F under physiological conditions.

Preliminary data suggest that p73, like p53, functions as a homotetramer¹¹. Previously, it was shown that overproduction of a p53 mutant (p53DD) encompassing the p53 oligomerization domain blocked p53 function in cells¹³. Furthermore, the p53 and p73 oligomerization domains do not bind to one another¹⁴. We therefore made an analogous p73 mutant (p73DD) and a point mutant derivative thereof (p73DD(L371P)) (Fig. 4a). The corresponding mutation in p53 (L344P) disrupts p53 oligomerization function^{15,16}.

As expected, p73DD bound to p73, and the closely related p63, but not p53 (Fig. 4b). p73DD blocked sequence-specific DNA binding by p73 *in vitro* (data not shown) and blocked p73-dependent transcriptional activation (Fig. 4c) and apoptosis (Fig. 4d) in tumour cells lacking p53. These effects were specific as they were not observed with either p73DD(L371P) (Fig. 4c, d) or p53DD (data not shown). Furthermore, p53-dependent transcriptional activation (Fig. 4c) and p53-dependent apoptosis (Fig. 4d) were not affected by p73DD.

Next, we transiently transfected SAOS2 cells to produce E2F-1 along with the cell-surface marker CD19. We monitored the number of CD19-positive cells undergoing apoptosis by propidium iodide staining and fluorescence-activated cell sorting (FACS). Co-production of p73DD, but not p73DD(L371P), with E2F-1 led to a

substantial decrease in the number of cells undergoing apoptosis (Fig. 5a). Similarly, p73DD blocked apoptosis in SAOS2 cells that were induced to express E2F-1 with doxycycline and in C33A cervical carcinoma cells transiently transfected to produce E2F-1 (data not shown). p73DD and p73DD(L371P) had no measurable effects on apoptosis in the absence of ectopically produced E2F-1 (Fig. 4d; and data not shown). These findings were corroborated by experiments in which apoptosis of the E2F-1-inducible SAOS2 cell line was monitored by immunofluorescence followed by DAPI (4',6-diamidino-2-phenylindole dihydrochloride) staining. Eight hours after E2F-1 induction more than 90% of cells that were transfected to produce p73DD appeared viable, whereas more than 90% of cells producing p73DD(L371P) appeared apoptotic (Fig. 5b; and data not shown). At much later time points, cells co-producing p73DD and E2F-1 eventually died, suggesting that p73DD does not completely block E2F-1-induced apoptosis (data not shown).

The availability of *p53*^{-/-} mice and *p73*^{-/-} mice allowed us to also test the requirement for p73 in E2F-1-mediated apoptosis in genetically defined cells. Wild-type mouse embryo fibroblasts (MEFs) underwent apoptosis, as measured by TUNEL (terminal deoxynucleotidyl transferase (TdT)-mediated dUTP nick end labelling) staining, after the production of E2F-1 by transient transfection (Fig. 5c). In comparison, apoptosis was substantially diminished in *p53*^{-/-} MEFs and in *p73*^{-/-} MEFs, and was virtually undetectable in *p53*^{-/-};*p73*^{-/-} MEFs (Fig. 5c). Similarly, *p53*^{-/-} MEFs were more susceptible to apoptosis than *p53*^{-/-};*p73*^{-/-} MEFs after introduction of E2F-1 by retroviral infection (data not shown). In this system, however, we observed lower levels of apoptosis in wild-type MEFs, which approximated the levels of apoptosis observed in *p73*^{-/-} MEFs. The reason for this is unclear, but may be due to technical differences between the two assays. Together, however, both systems show that p73 contributes to E2F-1-induced apoptosis in the absence of p53.

Many human tumours carry mutations that directly or indirectly inactivate the retinoblastoma protein (pRB) and thus de-repress E2F-responsive genes¹². The finding that p73 is induced by E2F was probably a factor both in the observation that p73 levels typically increase during neoplastic progression¹¹ and in the high levels of p73 observed in cells in which pRB is inactivated by SV40 T antigen or the adenoviral E1A protein¹⁷.

Induction of p73 provides a potential mechanism whereby E2F-1 can kill cells in the absence of p53. It is important to note, however, that production of E2F-1(1-374) in SAOS2 cells induces apoptosis^{1,2} and yet, as shown here, does not measurably induce p73. Both wild-type E2F-1 and E2F-1(1-374) have recently been shown to inhibit receptor mediated survival signals, such as activation of NF- κ B, through downregulation of TNF-receptor-associated factor 2 (TRAF2) expression¹⁸. This activity of E2F-1, which is clearly distinct from the ability to activate p73, might account for our finding that neither p73DD nor homozygous deletion of *p73* completely abrogated E2F-1-induced apoptosis.

E2F-1-induced apoptosis has been linked to the transcriptional activation of ARF and consequent stabilization of p53 (ref. 9). The fact that ARF itself does not induce apoptosis has led to the suggestion that E2F-1 sends a collateral signal that cooperates with p53 to induce death⁹. In light of our findings, induction of p73 might constitute such a signal. This would account for the diminished killing of *p73*^{-/-} MEFs by E2F-1 in transient transfection assays despite the presence of wild-type p53. Activation of p73 by E2F-1 is also required for apoptosis after T-cell receptor hyperstimulation¹⁹.

The fact that E2F-1 promotes both cell-cycle progression and apoptosis suggests that tumour cells must balance these two activities in a way that favours proliferation *in vivo*. SAOS2 cells, despite lacking wild-type pRB², produce low levels of both E2F-1 and p73. Nonetheless, our results indicate that superactivation of E2F-1 in these cells can activate p73 and induce apoptosis in the

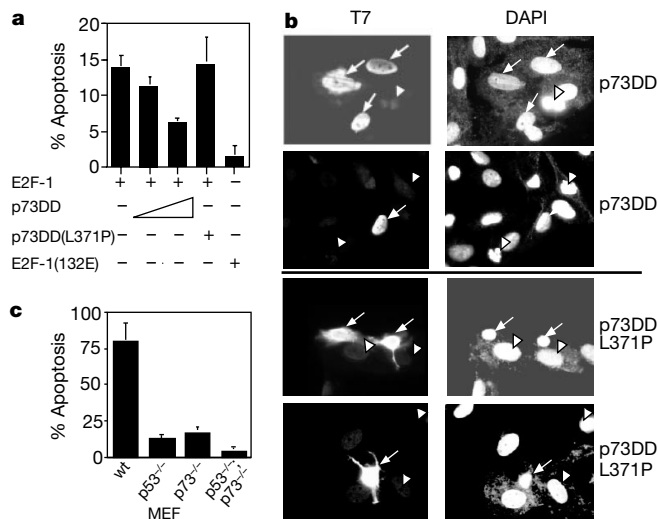


Figure 5 Inactivation of p73 inhibits E2F-1-induced apoptosis **a**, SAOS2 cells were transfected to produce wild-type E2F-1 or E2F-1(132E) in the presence or absence of increasing amounts of p73DD. The percentage of cells undergoing apoptosis, as measured by < 2N (subdiploid) DNA content, was determined by FACS 60 h later and corrected for the percentage of cells undergoing apoptosis after transfection with vector alone. Error bar indicates 1 standard error. **b**, SAOS2 cells with wild-type E2F-1 under the control of an inducible promoter were transfected to produce T7-epitope-tagged p73DD or p73DD(L371P). Apoptosis of transfected cells after E2F-1 induction was monitored by anti-T7 immunofluorescence and DAPI staining. Examples of transfected and untransfected cells are shown by arrows and arrowheads, respectively. Shown are representative fields 8 h after E2F-1 induction. Note that some untransfected cells appear viable because induction of apoptosis in this system is not entirely synchronous. **c**, Mouse embryo fibroblasts (MEFs) with the indicated genotypes were transfected to produce HA-tagged E2F-1. Shown is the percentage of HA-positive cells that were TUNEL-positive 24 h later. Error bar indicates 1 standard error ($n = 4$ independent experiments). Qualitatively similar results were seen at 48 h.

absence of p53. This finding may be relevant to efforts to treat human cancers using adenoviral vectors encoding E2F-1 (refs 8, 20) or with small molecule E2F-1 agonists²¹. □

Methods

Cell lines and mouse embryo fibroblasts

We grew SAOS2 subclones in which E2F-1 complementary DNAs were under the control of a doxycycline-inducible promoter as described¹⁰. We carried out SAOS2 transfections using the BBS/calcium phosphate method²².

We made *p53*^{-/-} embryos and *p73*^{-/-} embryos by crossing the respective heterozygous mice on a 129 background (*p73*^{+/-} mice were a gift from A. Yang and F. McKeon, Harvard University). *p53*^{+/-} mice were crossed to *p73*^{+/-} mice to create mice that were heterozygous for both *p53* and *p73* (*p53*^{+/-}; *p73*^{+/-}). We mated *p53*^{+/-}; *p73*^{+/-} mice to generate *p53*^{-/-}; *p73*^{-/-} embryos. We obtained embryos from pregnant females at day 13.5, and isolated MEFs as described²³. Genotypes were confirmed by PCR²⁴.

We plated 10⁵ early passage MEFs on coverslips and transfected them 24 h later with Lipofectamine Plus (Gibco-BRL) according to the manufacturer's instructions.

Plasmids

The p73 cDNA encoding amino acids 327–636 was PCR amplified with Pfu DNA polymerase (Stratagene) using primers introducing a 5' *Bam*HI site and a 3' *Eco*RI site, and ligated into pcDNA3–T7. pcDNA3–T7–p73DD(L371P) was made by site-directed mutagenesis and confirmed by DNA sequence analysis.

Plasmid pcDNA–HA–p53 has been described¹⁷. The pcDNA3–HA human p73 plasmids were a gift from G. Melino²⁵. p63α and p63γ cDNAs (a gift from S. Ikawa) were PCR amplified with Pfu DNA polymerase (Stratagene) using primers that introduced a 5' *Bam*HI site and a 3' *Not*I site, and ligated into pcDNA3–EE (W.G.K., unpublished data). The p21-promoter-luciferase plasmid (a gift from D. Cohen and K. Yu) has been described²². pRc–CMV–HA–E2F-1, pRc–CMV–HA–E2F2, pRc–CMV–HA–E2F3, pRc–CMV–HA–E2F4, pRc–CMV–HA–E2F-1(1–368) and pRc–CMV–E2F-1(E132) have been described²⁶.

We generated 857 bp of p73 promoter sequence upstream of the first nucleotide of p73 exon 1 (ref. 27) from the BAC clone, 190 O18 (Research Genetics), by mRS-PCR (multiplex restriction site PCR) using an anti-sense primer specific to exon 1 (5'-CCGTCCGACGCCCCGGGCA-3') and a mixture of restriction-site oligonucleotide primers²⁸. The amplified promoter fragment was cloned into the pGEM T-Easy vector (Promega; p73-pGEM T) and sequenced. The p73 fragment was then excised by digestion with *Nco*I, blunt-ended with Klenow polymerase, digested with *Sall*I, and subcloned into pGL3 (Promega) cut with *Sma*I and *Xho*I. The 5' deletion constructs were made by cutting the p73 promoter at unique *Pvu*II (–218) or *Not*I (–117) restriction sites followed by plasmid religation.

Immunoblot analysis

Immunoblot assays were performed essentially as described^{17,29} using 200–500 µg of nonidet P40 cell extract per lane with anti-p73 (ER15), anti-p63 (4A4) or E2F-1 (SC-193; Santa Cruz). Protein loading was normalized by Bradford assay.

Northern blot analysis

SAOS-2 cells (2 × 10⁶) inducibly expressing wild-type E2F-1 or the indicated E2F-1 mutants were plated 24 h before treatment with 2 µg ml⁻¹ doxycycline. At various time points after E2F-1 induction, we isolated RNA by standard techniques¹⁰ and analysed it by northern blotting using ³²P-dCTP-labelled p73 or p63 cDNA probes. The blots were then stripped and re-hybridized with a ³²P-dCTP-labelled GAPDH probe.

Luciferase reporter assays

SAOS2 cells were co-transfected with 0.5 µg of the indicated reporter plasmid, 0.5 µg pCMV–βgal and 1 µg of the indicated expression plasmid (p73β or p53) in the presence or absence of increasing amounts of pcDNA–p73DD (0.5–2 µg) or 2 µg of pcDNA–T7–p73DD(L371P). The total amount of DNA was normalized to 5.5 µg with pcDNA3 (Invitrogen). Luciferase and β-galactosidase assays were done 48 h after transfection.

In vitro translation and immunoprecipitation

We carried out *in vitro* transcription/translation using TNT reticulocyte lysate (Promega) according to the manufacturer's instructions. Immunoprecipitations were performed essentially as described¹⁷ in reactions containing 20 µl of translate and 1 µg of anti-T7 antibody (Novagen) in 500 µl of NETN (20 mM Tris pH 8, 100 mM NaCl, 1 mM EDTA, 0.5% nonidet-P40).

Flow cytometric apoptosis assays

SAOS2 cells were transfected with the indicated expression plasmids along with 3 µg of a plasmid encoding the cell-surface marker CD19. Apoptosis of CD19-positive cells was analysed 48–60 h after transfection by FACS as described²².

Immunofluorescence staining

We transfected E2F-1-inducible SAOS2 subclones with plasmids encoding T7–p73DD or

T7–p73DD(L371P). Twenty-four hours after transfection (8 h after removing the DNA precipitates), the cells were collected and replated on coverslips. Twenty-four hours later, doxycycline was added to induce E2F-1 production. At various time points thereafter, we stained cells with DAPI as well as with a mouse anti-T7 antibody (1:500 dilution; Novagen) using a rhodamine-conjugated anti-mouse antibody (1: 1000 dilution; Boehringer-Mannheim) as described²². MEF TUNEL assays were performed using a TUNEL Apoptosis Kit (Promega) according to the manufacturer's instructions after which the cells were stained with DAPI and mouse anti-HA antibody (1:50; 12CA5 hybridoma supernatant) as described above.

Received 7 July; accepted 31 August 2000.

- Phillips, A., Bates, S., Ryan, K., Helin, K. & Vousden, K. Induction of DNA synthesis and apoptosis are separable functions of E2F-1. *Genes Dev.* **11**, 1853–1863 (1997).
- Hsieh, J., Fredersdorf, S., Kouzarides, T., Martin, K. & Lu, X. E2F1-induced apoptosis requires DNA binding but not transactivation and is inhibited by the retinoblastoma protein through direct interaction. *Genes Dev.* **11**, 1840–1852 (1997).
- Wu, X. & Levine, A. J. p53 and E2F1 Cooperate to mediate apoptosis. *Proc. Natl Acad. Sci. USA* **91**, 3602–3606 (1994).
- Qin, X. Q., Livingston, D. M., Kaelin, W. G. & Adams, P. Deregulated E2F1 expression leads to S-phase entry and p53-mediated apoptosis. *Proc. Natl Acad. Sci. USA* **91**, 10918–10922 (1994).
- Holmberg, C., Helin, K., Sehested, M. & Karlstrom, O. E2F-1-Induced p53-independent apoptosis in transgenic mice. *Oncogene* **17**, 143–155 (1998).
- Nip, J. *et al.* E2F-1 cooperates with topoisomerase II inhibition and DNA damage to selectively augment p53-independent apoptosis. *Mol. Cell. Biol.* **17**, 1049–1056 (1997).
- Agar, R. *et al.* Adenoviral delivery of E2F-1 directs cell cycle reentry and p53-independent apoptosis in postmitotic adult myocardium in vivo. *J. Clin. Invest.* **100**, 2722–2728 (1997).
- Hunt, K. *et al.* Adenovirus-mediated overexpression of the transcription factor E2F-1 induces apoptosis in human breast and ovarian carcinoma cell lines and does not require p53. *Cancer Res.* **57**, 4722–4726 (1997).
- Sherr, C. Tumor surveillance via the ARF–p53 pathway. *Genes Dev.* **12**, 2984–2991 (1998).
- Bates, S. *et al.* p14ARF links the tumour suppressors RB and p53. *Nature* **395**, 124–125 (1998).
- Kaelin, W. G. The emerging p53 gene family. *J. Natl Cancer Institute* **91**, 594–598 (1999).
- Dyson, N. The regulation of E2F by pRB-family proteins. *Genes Dev.* **12**, 2245–2262 (1998).
- Shaulian, E., Zauberman, A., Ginsberg, D. & Oren, M. Identification of a minimal transforming domain of p53: negative dominance through abrogation of sequence-specific DNA binding. *Mol. Cell. Biol.* **12**, 5581–5592 (1992).
- Davison, T. *et al.* p73 and p63 are homotetramers capable of weak heterotypic interactions with each other but not with p53. *J. Biol. Chem.* **274**, 18709–18714 (1999).
- Lomax, M., Barnes, D., Hupp, T., Pickles, S. & Campjohn, R. Characterization of p53 oligomerization domain mutations isolated from Li-Fraumeni and Li-Fraumeni like family members. *Oncogene* **17**, 643–649 (1998).
- Davison, T., Yin, P., Nie, E., Kay, C. & Arrowsmith, C. Characterization of the oligomerization defects of two p53 mutants found in families with Li-Fraumeni and Li-Fraumeni syndrome. *Oncogene* **17**, 651–656 (1998).
- Marin, M. C. *et al.* Viral oncoproteins discriminate between p53 and the p53 homologue p73. *Mol. Cell. Biol.* **18**, 6316–6324 (1998).
- Phillips, A., Ernst, M., Bates, S., Rice, N. & Vousden, K. E2F-1 potentiates cell death by blocking antiapoptotic signaling pathways. *Mol. Cell* **4**, 771–781 (1999).
- Lissy, N. A., Davis, P. K., Irwin, M., Kaelin, W. G. & Dowdy, S. F. A common E2F-1 and p73 pathway mediates cell death induced by TCR activation. *Nature* **407**, 642–645 (2000).
- Fuoy, J. *et al.* Overexpression of E2F-1 in glioma triggers apoptosis and suppresses tumour growth in vitro and in vivo. *Nature Med.* **4**, 685–690 (1998).
- Chen, Y. *et al.* Selective killing of transformed cells by cyclin/cyclin-dependent kinase 2 antagonists. *Proc. Natl Acad. Sci. USA* **96**, 4325–4329 (1999).
- Jost, C., Marin, M. & Kaelin, W. G. p73 is a human p53-related protein that can induce apoptosis. *Nature* **389**, 191–194 (1997).
- Brugarolas, J., Bronson, R. & Jacks, T. p21 is a critical CDK2 regulator essential for proliferation control in Rb-deficient cells. *J. Cell Biol.* **141**, 503–514 (1998).
- Yang, A. *et al.* p73-deficient mice have neurological, phenomonal and inflammatory defects but lack spontaneous tumors. *Nature* **404**, 99–103 (2000).
- De Laurenzi, V. *et al.* Two new p73 splice variants, γ and δ, with different transcriptional activity. *J. Exp. Med.* **188**, 1763–1768 (1998).
- Krek, W. *et al.* Negative regulation of the growth-promoting transcription factor E2F-1 by a stably bound cyclin A-dependent protein kinase. *Cell* **78**, 1–20 (1994).
- Kaghad, M. *et al.* Monoallelically Expressed Gene Related to p53 at 1p36, a region frequently deleted in neuroblastoma and other human cancers. *Cell* **90**, 809–819 (1997).
- Weber, K. L., Bolander, M. E. & Sarkar, G. Rapid acquisition of unknown DNA sequence adjacent to a known segment by multiplex restriction site PCR. *BioTechniques* **25**, 415–419 (1998).
- Yang, A. *et al.* p63, a p53 homologue at 3q27–29, encodes multiple products with transactivating, death inducing, and dominant-negative activities. *Mol. Cell* **2**, 305–316 (1998).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy form the London editorial office of Nature.

Acknowledgements

We thank A. Yang and F. McKeon for *p73*^{+/-} mice and K. Baker for help with the apoptosis assays. M.I. is an American Cancer Society postdoctoral fellow. W.G.K. is a Howard Hughes Medical Institute assistant investigator.

Correspondence and requests for material should be addressed to W.G.K. (e-mail: william_kaelin@dfci.harvard.edu).

S-RNase uptake by compatible pollen tubes in gametophytic self-incompatibility

Doan-Trung Luu, Xike Qin, David Morse & Mario Cappadocia

Biology Department, University of Montreal, Montreal, Quebec H1X 2B2, Canada

Many flowering plants avoid inbreeding through a genetic mechanism termed self-incompatibility. An extremely polymorphic S-locus¹ controls the gametophytic self-incompatibility system that causes pollen rejection (that is, active arrest of pollen tube growth inside the style) when an S-allele carried by haploid pollen matches one of the S-alleles present in the diploid style. The only known product of the S-locus is an S-RNase expressed in the mature style². The pollen component to this cell–cell recognition system is unknown and current models^{3,4} propose that it either acts as a gatekeeper allowing only its cognate S-RNase to enter the pollen tube, or as an inhibitor of non-cognate S-RNases. In the latter case, all S-RNases are presumed to enter pollen tubes; thus, the two models make diametrically opposed predictions concerning the entry of S-RNases into compatible pollen. Here we use immunocytochemical labelling of pollen tubes growing in styles to show accumulation of an S-RNase in the cytoplasm of all pollen-tube haplotypes, thus providing experimental support for the inhibitor model.

We have used a monospecific polyclonal antibody directed against a specific S-RNase to label pollen tubes growing in incompatible styles. S-RNases contain an RNase activity domain required for pollen rejection^{5,6} and a hypervariable region involved in allele-specific recognition^{7,8}. The 15-amino-acid peptide used to prepare the antibody corresponded to the hypervariable region of the S₁₁-RNase, and the antibody recognizes the S₁₁-RNase but does not recognize the S₁₃-RNase, which differs from the former by only 3 amino acids in this region⁸. In self-pollinated styles of plant V22 (genotype S₁₁S₁₃; expressing both S₁₁ and S₁₃ RNases), the arrest of pollen-tube growth occurs between the upper third and upper half of the style and is complete by 48 hours post-pollination. Styles collected 18 hours post-pollination (Fig. 1a), which contain still-growing pollen tubes, were thus selected for the immunolabelling. Light microscopy was used to identify the stylar regions containing the apical tips of pollen tubes, where elongation occurs, and transverse sections from these regions were then stained for transmission electron microscopy (TEM) with the anti-S₁₁ and gold-labelled secondary antibodies. Pollen tubes appear as small, irregularly shaped cells amid the larger, rounder cells of the transmitting tissue, and their identity was determined by comparing the morphology of unpollinated and pollinated styles^{9,10} and by *in situ* hybridization with a pollen-specific gene probe (data not shown). In these samples (not stained with osmium, in order to preserve their antigenicity), apical regions of pollen tubes contain darker-staining cytoplasm at their peripheral regions surrounding either a paler electron-lucent vacuole or numerous electron-lucent secretory vesicles, depending on the position of the section relative to the pollen tip¹¹.

Pollen-tube labelling (Fig. 1b and c) occurs principally in the cytoplasm. More distal regions of pollen tubes (not shown) appear collapsed, devoid of cytoplasm, and are not labelled. In contrast to the pollen tubes, cells of the transmitting tissue generally show labelling over the lighter-staining regions. The presence of S-RNases inside membrane-bounded compartments is expected for a secreted protein and is consistent with the label found in the extracellular matrix (Fig. 1b and c). For the experiments shown in Fig. 1, label

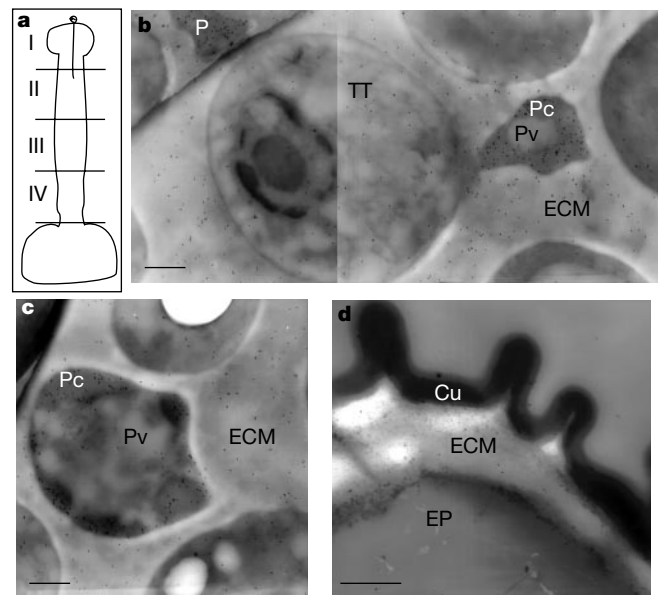


Figure 1 S₁₁-RNase entry into incompatible pollen tubes. **a**, Self-pollinated styles of an S₁₁S₁₃ plant were fixed 18 hours post-pollination, sectioned in region II, and stained sequentially with the anti-S₁₁ antibody and a 20-nm gold-labelled secondary antibody. **b, c**, The apical tips of all growing pollen tubes (P) are labelled to a greater extent than either the transmitting tissue cells (TT) or the extracellular matrix (ECM) separating the cells. Label inside pollen tubes appears associated with the cytoplasm (Pc) rather than the vacuolar spaces (Pv), unlike the labelling found in transmitting tissue cells. **d**, Only background labelling is present in the cytoplasm, the surrounding cell wall or the cuticle (Cu) of epidermal cells (EP), which do not express S-RNases. All scale bars are 1 μ m.

density in the extracellular matrix (6 ± 2 gold particles per μm^2) is twice that found inside the transmitting tissue cells (3 ± 0.6 gold particles per μm^2) and substantially less than that inside the pollen tube cytoplasm (32 ± 10 gold particles per μm^2). All label densities are greater than the background (0.5 ± 0.3 gold particles per μm^2), measured either in the absence of the primary antibody (not shown) or in stylar epidermal tissues stained with the anti-S₁₁ antibody (Fig. 1d). Epidermal tissues are known not to express S-RNases^{12,13}, and are not expected to cross-react with the antibody.

To confirm genotype-independent S-RNase uptake by pollen tubes, we labelled sections from V22 styles following a fully compatible cross. Pollen was obtained from the S-homozygous plant VF60, so that all have an identical S₁₂-allele constitution. All pollen tubes are stained with the anti-S₁₁ antibody, and the pattern of label accumulation inside the cytoplasm (Fig. 2a) is indistinguishable from that found in fully incompatible crosses (Fig. 1). The label density inside S₁₂-pollen tubes (31 ± 15 gold particles per μm^2) is higher than that found in the extracellular matrix (3.8 ± 0.8 gold particles per μm^2) or inside the transmitting tissue cells (2.6 ± 0.8 gold particles per μm^2). Thus, the S₁₁-RNase enters S₁₂ pollen tubes. In another experiment, styles of V22 were pollinated with S₁₃ and S₁₄ pollen (a semi-compatible cross). Once again, the S₁₁-RNase was found in the cytoplasm of all pollen tubes (Fig. 2b) even though none of the pollen was S₁₁. Despite a lower overall intensity of label, the label density in pollen cytoplasm (12 ± 4 gold particles per μm^2) was again fivefold higher than that found in the extracellular matrix (2.2 ± 0.9 gold particles per μm^2) and tenfold higher than the transmitting tissue cells (1.4 ± 0.7 gold particles per μm^2). As a last control for the specificity of the reaction, sections containing compatible S₁₁ and S₁₃ pollen tubes growing in S₁₂S₁₄ styles were stained with the anti-S₁₁ antibody (Fig. 2c). As expected, given the absence of an S₁₁-RNase, only background label density was measured (0.6 ± 0.3 gold particles per μm^2). We note that the lack of label in the pollen tube cytoplasm indicates that the antibody does not cross-react with a pollen-specific antigen. Taken

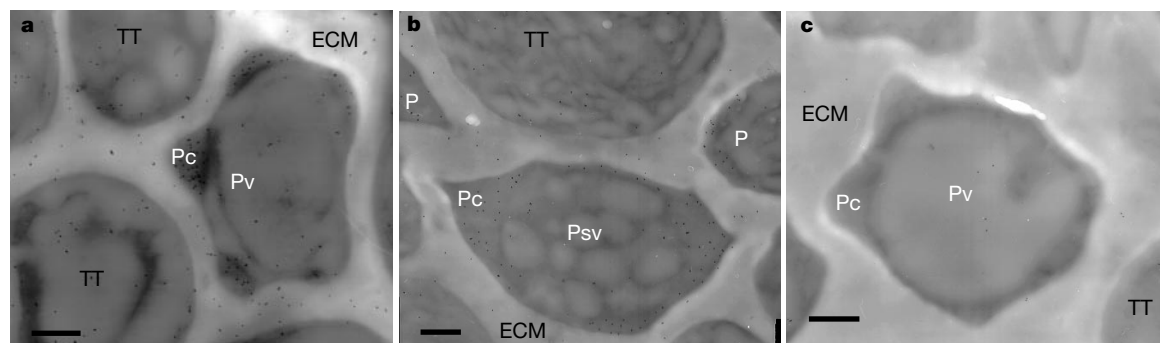


Figure 2 S_{11} -RNase entry into compatible pollen tubes. Pollinated styles were treated as described in Fig. 1. **a**, A fully compatible cross of $S_{11}S_{13}$ styles pollinated with S_{12} pollen. **b**, A semi-compatible cross of $S_{11}S_{13}$ plants pollinated with S_{13} and S_{14} pollen. The region observed is close to the middle of the style and is deduced to contain only fast-growing

compatible S_{14} pollen tubes. The several electron-lucent regions correspond to secretory vesicles (Psv). **c**, A fully compatible cross of $S_{12}S_{14}$ styles pollinated with S_{11} and S_{13} pollen, showing the background labelling of the pollen tubes in the absence of an S_{11} -RNase. Cells are labelled as in Fig. 1 and all scale bars are 1 μ m.

together, these results show an active uptake or accumulation of the S_{11} -RNase in pollen tube cytoplasm. We also note that, as all pollen tubes are similarly labelled, S-RNase accumulation in pollen tubes is genotype-independent. All these observations are inconsistent with a gatekeeper model requiring specific entry of an S-RNase only into pollen tubes carrying the corresponding S-allele.

To place in perspective our findings showing S-RNase uptake in both compatible and incompatible pollen tubes growing *in vivo*, we note that genotype-independent accumulation of S-RNases in pollen tubes germinated *in vitro* in solutions of purified S-RNases has been previously reported¹⁴. However, in these studies, both pollen-tube RNA degradation¹⁴ and growth arrest¹⁵ were genotype-independent, and thus the genotype-independent S-RNase uptake was attributed to an artefact of *in vitro* germinated pollen¹⁶.

Our studies support a model for self-incompatibility in which an RNase inhibitor must be present inside pollen tubes. The current inhibitor model of gametophytic self-incompatibility³ identifies pollen-S with this RNase inhibitor. Pollen-S, like stylar S-RNase, would contain two domains, one binding in an allele-specific way to the recognition domain of its corresponding S-RNase, the other binding to (and inhibiting) the activity domain of any other S-RNase. For the self-incompatibility system to work, one must assume that binding to the two domains is mutually exclusive, and that binding to the recognition domain is thermodynamically favoured over binding to the RNase activity domain. Thus, both binding of pollen-S to the recognition domain (for incompatible crosses) or to the RNase activity domain (for compatible crosses) would result in S-RNase accumulation in pollen tubes, as we show here.

How does the inhibitor model deal with the breakdown of self-incompatibility in pollen-part mutations? In most cases^{1,17} the breakdown has been associated with the presence of two different S-alleles in the same pollen grain, a situation similar to the compatibility of tetraploids derived from self-incompatible diploids (also termed competitive interaction). To explain the competitive interaction, the inhibitor model requires a mechanism enabling pollen-S to bind to the RNase activity domain even in the presence of its cognate S-RNase recognition domain. In other cases, where competitive interaction had been ruled out^{18,19}, pollen compatibility was presumed to result from deletion or inactivation of the pollen S-allele. However, the inhibitor model for pollen-S predicts that deletions of pollen-S should be lethal¹⁷. To reconcile the inhibitor model with the compatibility of pollen-S deletions, we propose that the two functions of RNase inhibition and allele-specific recognition are carried by separate proteins, RNase inhibitor and pollen-S, respectively. In our view, all pollen contain a general RNase inhibitor that can potentially bind and inactivate any S-RNase. This binding can only be prevented when pollen-S binds to the recognition domain of its cognate S-RNase, which would lead to the self-incompatibility reaction by activation of the RNase activity.

Our version of the inhibitor model predicts that deletion of pollen-S would produce fully compatible pollen, and could best be confirmed by recovery of self-compatible pollen-part mutants generated by mutagenesis of S-homozygous plants. The compatibility of S-heteroallelic pollen could now be due either to a mechanism impeding pollen-S binding to the recognition domain of its cognate S-RNase or to a simple lack of pollen-S expression. □

Methods

Plant material

Two fully self-incompatible but cross-compatible diploid *Solanum chacoense* Bitt. lines were obtained from the Potato Introduction Station (Sturgeon Bay, Wisconsin). The self-incompatibility alleles in the parental line PI458314 were identified by genetic crosses and gene sequence as S_{11} and S_{12} (ref. 20) and the alleles in the parental line PI230582 identified as S_{13} and S_{14} (ref. 21). The F_1 hybrids used in the present study were obtained by genetic crosses of the parental lines, and include G4 ($S_{12}S_{14}$) (ref. 22) and V22 ($S_{11}S_{13}$) (ref. 23), whereas VF60 ($S_{12}S_{12}$) was a selfed line from the parental line PI458314 (ref. 24).

Microscopy and immunolabelling

Styles from pollinated flowers were harvested 18 hours post-pollination and either fixed immediately in 0.5% glutaraldehyde in 0.1 M PBS (pH 7.4) as preparatory for immunogold labelling²⁵, or squashed and stained with a 0.2% solution of aniline blue to estimate the location of the pollen tube tips by ultraviolet microscopy²⁶. Samples fixed with glutaraldehyde for 1.5 hours at room temperature were dehydrated and embedded in LR White (Electron Microscopy Services). Regions of the blocks containing the tips of growing pollen tubes were selected by observation of sections stained with toluidine blue. Regions behind the growing pollen tips do not contain cytoplasm and are not stained with the antibody. The selected regions were then thin-sectioned (150–170 nm 'purple' sections) and placed on Formvar-coated grids. Sections were incubated with a 1/50 dilution of a rabbit anti- S_{11} antibody for 1 hour at room temperature, washed with PBST (PBS containing 0.05% w/v Tween 20) then incubated with a secondary goat anti-rabbit conjugated to 20-nm gold particles (Ted Pella Inc.). To confirm identification, some samples were hybridized to an antisense pollen-specific *Solanum* gene probe homologous to the *Petunia hybrida* PGP35 (Accession number AF161330). The single-strand-RNA probe was prepared using T7 RNA polymerase and Digoxigenin (DIG)-labelled UTP (Boehringer Mannheim). The probe was hybridized to the sections for 2 hours at 42 °C in a solution containing 40% formamide as described²⁷. The DIG label was detected using a sheep anti-DIG conjugated to 10-nm gold particles (Ted Pella Inc.). The sequence of antibodies for double labelling subsequent to the *in situ* hybridization was either (1) 10-nm anti-DIG, anti- S_{11} , 20-nm anti-rabbit or (2) anti- S_{11} followed by a mixture of 10-nm anti-DIG and 20-nm anti-rabbit. Results were identical with both protocols. Sections were observed using a JEOL JEM 100S operating at 80 kV. The rabbit anti- S_{11} antibody was prepared by Cocalico Biological Inc. against the peptide KPRLTYNYFSDKMLN synthesized as a branched multiple-antigen peptide (Research Genetics). On standard western blots, the antibody reacts only with the S_{11} -RNase⁸.

Received 5 June; accepted 19 July 2000.

- de Nettancourt, D. *Incompatibility in Angiosperms* (Springer, New York, 1977).
- McClure, B. *et al.* Style self-incompatibility products of *Nicotiana glauca* are ribonucleases. *Nature* **32**, 955–957 (1989).
- Kao, T.-H. & McClure, B. How flowering plants discriminate between self and non-self pollen to prevent inbreeding. *Proc. Natl Acad. Sci. USA* **93**, 12059–12065 (1996).
- Dodds, P., Clarke, A. & Newbigin, E. A molecular perspective on pollination in flowering plants. *Cell* **85**, 141–144 (1996).
- Huang, S., Lee, H.-S., Karunanandaa, B. & Kao, T.-H. Ribonuclease activity of *Petunia inflata* S proteins is essential for rejection of self pollen. *Plant Cell* **6**, 1021–1028 (1994).
- McClure, B., Gray, J., Anderson, M. & Clarke, A. Self-incompatibility in *Nicotiana glauca* involves

- degradation of pollen RNA. *Nature* **347**, 757–760 (1990).
7. Matton, D. P. *et al.* Hypervariable domains of self-incompatibility RNases mediate allele-specific pollen recognition. *Plant Cell* **9**, 1757–1766 (1997).
 8. Matton, D. P. *et al.* The production of an S-RNase with dual specificity suggests a novel hypothesis for the generation of new S-alleles. *Plant Cell* **11**, 2087–2097 (1999).
 9. Cheung, A. The pollen tube growth pathway: its molecular and biochemical contributions and responses to pollination. *Sex. Plant Reprod.* **9**, 330–336 (1996).
 10. de Nettancourt, D. *et al.* Ultrastructural aspects of the self-incompatibility mechanism in *Lycopersicon peruvianum* Mill. *J. Cell Sci.* **12**, 403–419 (1973).
 11. Steer, M. W. & Steer, J. M. Pollen tube tip growth. *New Phytol.* **111**, 323–358 (1989).
 12. Anderson, M. *et al.* Cloning of cDNA for a stylar glycoprotein associated with expression of self-incompatibility in *Nicotiana glauca*. *Nature* **321**, 38–44 (1986).
 13. Certal, A. *et al.* S-RNases in apple are expressed in the pistil along the pollen tube growth path. *Sex. Plant Reprod.* **12**, 94–98 (1999).
 14. Gray, J., McClure, B., Böning, I., Anderson, M. & Clarke, A. Action of the style product of the self-incompatibility gene of *Nicotiana glauca* (S-RNase) on *in vitro* grown pollen tubes. *Plant Cell* **3**, 271–283 (1991).
 15. Jahnen, W., Lush, W. & Clarke, A. Inhibition of *in vitro* pollen tube growth by isolated S-glycoproteins of *Nicotiana glauca*. *Plant Cell* **1**, 501–510 (1989).
 16. Hess, D., Gresshoff, P., Fiehlitz, U. & Gleiss, D. Uptake of protein and bacteriophage into swelling and germinating pollen of *Petunia hybrida*. *Z. Pflanzenphysiol.* **74**, 371–376 (1974).
 17. Golz, J., Clarke, A. E. & Newbegin, E. Mutational approaches to the study of self-incompatibility: revisiting the pollen-part mutants. *Ann. Botany* **85**, 95–103 (2000).
 18. Lewis, D. Chromosome fragments and mutation of the incompatibility gene. *Nature* **190**, 990–991 (1961).
 19. Pandey, K. Elements of the S-gene complex. II. Mutations and complementation at the SI locus in *Nicotiana glauca*. *Heredity* **22**, 255–284 (1967).
 20. Saba-El-Leil, M., Rivard, S., Morse, D. & Cappadocia, M. The S11 and S13 self-incompatibility alleles in *Solanum chacoense* Bitt. are remarkably similar. *Plant Mol. Biol.* **24**, 571–583 (1994).
 21. Despres, C., Saba-El-Leil, M., Rivard, S., Morse, D. & Cappadocia, M. Molecular cloning of two *Solanum chacoense* S-alleles and a hypothesis concerning their evolution. *Sex. Plant Reprod.* **7**, 169–176 (1994).
 22. Van Sint Jan, V., Laublin, G., Birhman, R. & Cappadocia, M. Genetic analysis of leaf explant regenerability in *Solanum chacoense*. *Plant Cell Tissue Org. Cult.* **47**, 9–13 (1996).
 23. Veronneau, H., Lavoie, G. & Cappadocia, M. Genetic analysis of anther and leaf disk culture in two clones of *Solanum chacoense* Bitt. and their reciprocal hybrids. *Plant Cell Tissue Org. Cult.* **30**, 199–209 (1992).
 24. Rivard, S., Saba-El-Leil, M., Landry, B. & Cappadocia, M. RFLP analyses and segregation of molecular markers in plants produced by *in vitro* anther culture, selfing, and reciprocal crosses of two lines of self-incompatible *Solanum chacoense*. *Genome* **37**, 775–783 (1994).
 25. Vandenbosch, K. in *Electron Microscopy of Plant Cells* (eds Hall, J. & Hawes, C.) 181–218 (Academic, New York, 1991).
 26. Martin, F. Staining and observing pollen tubes in the style by means of fluorescence. *Stain Technol.* **34**, 125–128 (1959).
 27. McFadden, G. in *Electron Microscopy of Plant Cells* (eds Hall, J. & Hawes, C.) 220–255 (Academic, New York, 1991).

Acknowledgements

We thank L. Pelletier and N. Nassoury for technical assistance, G. Teodorescu for plant care, and S. McCormick, A. Cheung, T.-H. Kao and V. de Luca for helpful discussions. The work was supported by a fellowship from Programme Québécois de Bourses d'Excellence, Québec (D.-T.L.) and by grants from NSERC (M.C.) and Fonds pour la Formation de Chercheurs et l'Aide à la Recherche (D.M., M.C.).

Correspondence and requests for materials should be addressed to M.C. (e-mail: mario.cappadocia@umontreal.ca).

The large-scale organization of metabolic networks

H. Jeong^{*}, B. Tombor[†], R. Albert^{*}, Z. N. Oltvai[†] & A.-L. Barabási^{*}

^{*} Department of Physics, University of Notre Dame, Notre Dame, Indiana 46556, USA

[†] Department of Pathology, Northwestern University Medical School, Chicago, Illinois 60611, USA

In a cell or microorganism, the processes that generate mass, energy, information transfer and cell-fate specification are seamlessly integrated through a complex network of cellular constituents and reactions¹. However, despite the key role of these networks in sustaining cellular functions, their large-scale structure is essentially unknown. Here we present a systematic comparative mathematical analysis of the metabolic networks of

43 organisms representing all three domains of life. We show that, despite significant variation in their individual constituents and pathways, these metabolic networks have the same topological scaling properties and show striking similarities to the inherent organization of complex non-biological systems². This may indicate that metabolic organization is not only identical for all living organisms, but also complies with the design principles of robust and error-tolerant scale-free networks^{3–5}, and may represent a common blueprint for the large-scale organization of interactions among all cellular constituents.

An important goal in biology is to uncover the fundamental design principles that provide the common underlying structure and function in all cells and microorganisms^{6–13}. For example, it is increasingly appreciated that the robustness of various cellular processes is rooted in the dynamic interactions among its many constituents^{14–16}, such as proteins, DNA, RNA and small molecules. Scientific developments have improved our ability to identify the design principles that integrate these interactions into a complex system. Large-scale sequencing projects have not only provided complete sequence information for a number of genomes, but also allowed the development of integrated pathway–genome databases^{17–19} that provide organism-specific connectivity maps of metabolic and, to a lesser extent, other cellular networks. However, owing to the large number and diversity of the constituents and reactions that form such networks, these maps are extremely complex, offering only limited insight into the organizational principles of these systems. Our ability to address in quantitative terms the structure of these cellular networks has benefited from advances in understanding the generic properties of complex networks².

Until recently, complex networks have been modelled using the classical random network theory introduced by Erdős and Rényi^{20,21}. The Erdős–Rényi model assumes that each pair of nodes (that is, constituents) in the network is connected randomly with probability p , leading to a statistically homogeneous network in which, despite the fundamental randomness of the model, most nodes have the same number of links, $\langle k \rangle$ (Fig. 1a). In particular, the connectivity follows a Poisson distribution that peaks strongly at $\langle k \rangle$ (Fig. 1b), implying that the probability of finding a highly connected node decays exponentially ($P(k) \approx e^{-k}$ for $k \gg \langle k \rangle$). On the other hand, empirical studies on the structure of the World-Wide Web²², Internet²³ and social networks² have reported serious deviations from this random structure, showing that these systems are described by scale-free networks² (Fig. 1c), for which $P(k)$ follows a power-law, $P(k) \approx k^{-\gamma}$ (Fig. 1d). Unlike exponential networks, scale-free networks are extremely heterogeneous, their topology being dominated by a few highly connected nodes (hubs) which link the rest of the less connected nodes to the system (Fig. 1c). As the distinction between scale-free and exponential networks emerges as a result of simple dynamical principles^{24,25}, understanding the large-scale structure of cellular networks can not only provide valuable and perhaps universal structural information, but could also lead to a better understanding of the dynamical processes that generated these networks. In this respect the emergence of power-law distribution is intimately linked to the growth of the network in which new nodes are preferentially attached to already established nodes², a property that is also thought to characterize the evolution of biological systems¹.

To begin to address the large-scale structural organization of cellular networks, we have examined the topological properties of the core metabolic network of 43 different organisms based on data deposited in the WIT database¹⁹. This integrated pathway–genome database predicts the existence of a given metabolic pathway on the basis of the annotated genome of an organism combined with firmly established data from the biochemical literature. As 18 of the 43 genomes deposited in the database are not yet fully sequenced, and a substantial portion of the identified open reading frames are

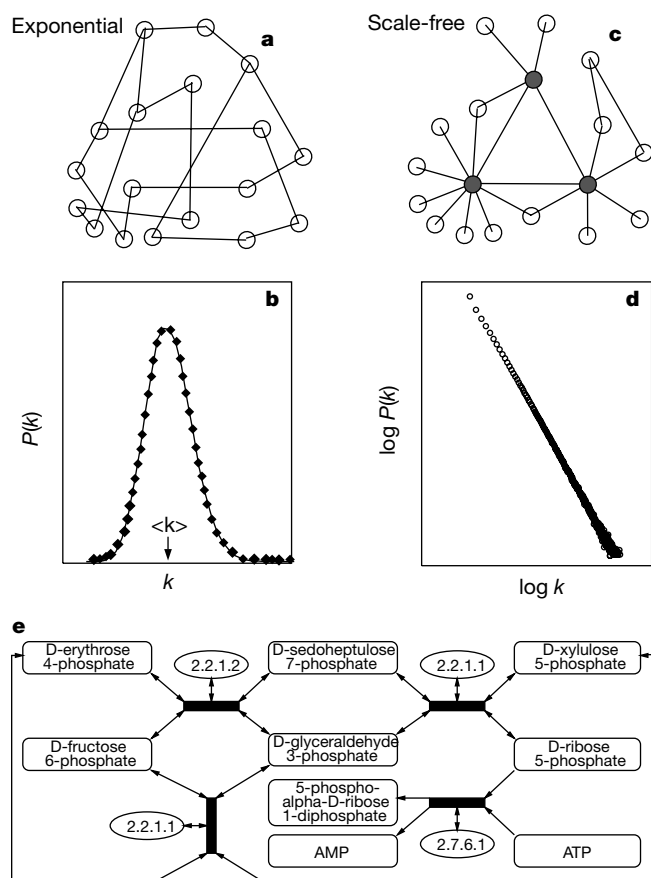


Figure 1 Attributes of generic network structures. **a**, Representative structure of the network generated by the Erdős-Rényi network model^{20,21}. **b**, The network connectivity can be characterized by the probability, $P(k)$, that a node has k links. For a random network $P(k)$ peaks strongly at $k = \langle k \rangle$ and decays exponentially for large k (that is, $P(k) \approx e^{-k}$ for $k \gg \langle k \rangle$ and $k \ll \langle k \rangle$). **c**, In the scale-free network most nodes have only a few links, but a few nodes, called hubs (red), have a very large number of links. **d**, $P(k)$ for a scale-free network has no well-defined peak, and for large k it decays as a power-law, $P(k) \approx k^{-\gamma}$, appearing as a straight line with slope $-\gamma$ on a log-log plot. **e**, A portion of the WIT database for *E. coli*. Each substrate can be represented as a node of the graph, linked through temporary educt–educt complexes (black boxes) from which the products emerge as new nodes (substrates). The enzymes, which provide the catalytic scaffolds for the reactions, are shown by their EC numbers.

functionally unassigned, the list of enzymes, and consequently the list of substrates and reactions (see Table 1 in Supplementary Information), will certainly be expanded in the future. Nevertheless, this publicly available database represents our best approximation for the metabolic pathways in 43 organisms and provides sufficient data for their unambiguous statistical analysis (see Methods and Supplementary Information).

As we show in Fig. 1e, we first established a graph theoretic representation of the biochemical reactions taking place in a given metabolic network. In this representation, a metabolic network is built up of nodes, the substrates, that are connected to one another through links, which are the actual metabolic reactions. The physical entity of the link is the temporary educt–educt complex itself, in which enzymes provide the catalytic scaffolds for the reactions yielding products, which in turn can become educts for subsequent reactions. This representation allows us systematically to investigate and quantify the topologic properties of various metabolic networks using the tools of graph theory and statistical mechanics²¹.

Our first goal was to identify the structure of the metabolic networks: that is, to establish whether their topology is best

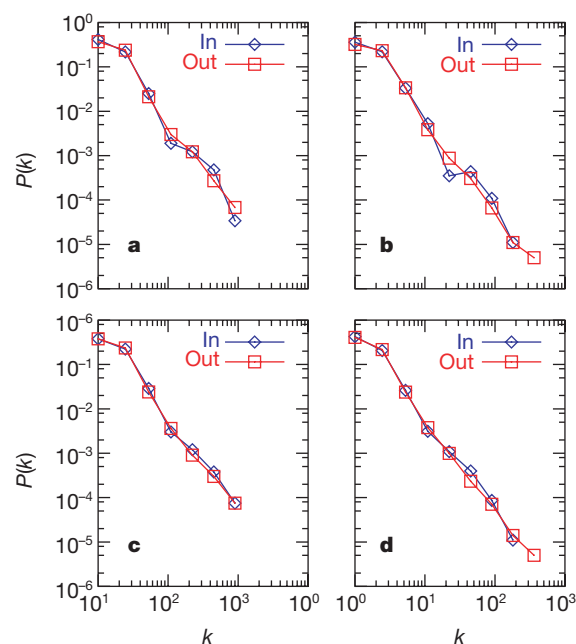


Figure 2 Connectivity distributions $P(k)$ for substrates. **a**, *Archaeoglobus fulgidus* (archae); **b**, *E. coli* (bacterium); **c**, *Caenorhabditis elegans* (eukaryote), shown on a log-log plot, counting separately the incoming (In) and outgoing links (Out) for each substrate. k_{in} (k_{out}) corresponds to the number of reactions in which a substrate participates as a product (educt). The characteristics of the three organisms shown in **a–c** and the exponents γ_{in} and γ_{out} for all organisms are given in Table 1 of the Supplementary Information. **d**, The connectivity distribution averaged over all 43 organisms.

described by the inherently random and uniform exponential model²¹ (Fig. 1a, b), or the highly heterogeneous scale-free model² (Fig. 1c, d). As illustrated in Fig. 2, our results convincingly indicate that the probability that a given substrate participates in k reactions follows a power-law distribution; in other words, metabolic networks belong to the class of scale-free networks. As under physiological conditions a large number of biochemical reactions (links) in a metabolic network are preferentially catalysed in one direction (the links are directed), for each node we distinguish between incoming and outgoing links (Fig. 1e). For instance, in *Escherichia coli* the probability that a substrate participates as an educt in k metabolic reactions follows $P(k) \approx k^{-\gamma_{in}}$, with $\gamma_{in} = 2.2$, and the probability that a given substrate is produced by k different metabolic reactions follows a similar distribution, with $\gamma_{out} = 2.2$ (Fig. 2b). We find that scale-free networks describe the metabolic networks in all organisms in all three domains of life (Fig. 2a–c; see Supplementary Information, also available at www.nd.edu/~networks/cell), indicating the generic nature of this structural organization (Fig. 2d).

A general feature of many complex networks is their small-world character²⁶, meaning that any two nodes in the system can be connected by relatively short paths along existing links. In metabolic networks these paths correspond to the biochemical pathway connecting two substrates (Fig. 3a). The degree of interconnectivity of a metabolic network can be characterized by the network diameter, defined as the shortest biochemical pathway averaged over all pairs of substrates. For all non-biological networks examined, the average connectivity of a node is fixed, which implies that the diameter of a network increases logarithmically with the addition of new nodes^{2,26,27}. For metabolic networks this implies that a more complex bacterium with more enzymes and substrates, such as *E. coli*, would have a larger diameter than a simple bacterium, such as *Mycoplasma genitalium*. We find, however, that the diameter of the metabolic network is the same for all 43 organisms,

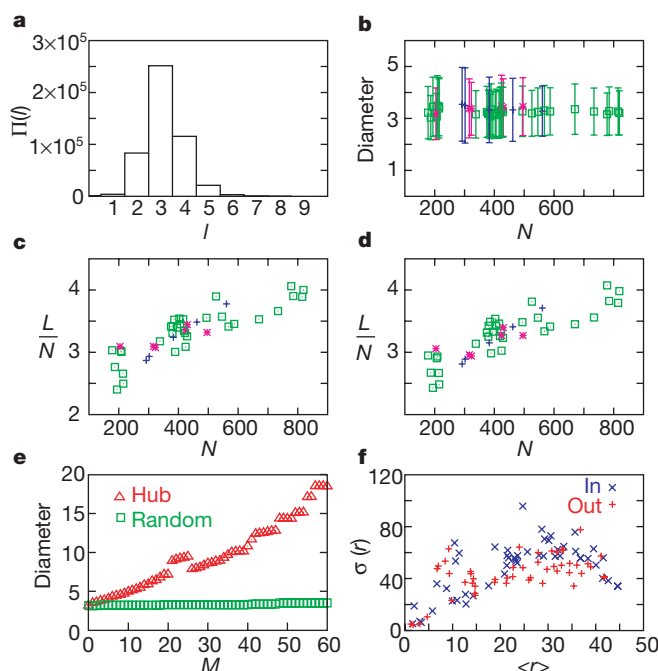


Figure 3 Properties of metabolic networks. **a**, The histogram of the biochemical pathway lengths, l , in *E. coli*. **b**, The average path length (diameter) for each of the 43 organisms. Error bars represent standard deviation $\sigma \approx \langle l^2 \rangle - \langle l \rangle^2$ as determined from $\Pi(l)$ (shown in **a** for *E. coli*). **c**, **d**, Average number of incoming links (**c**) or outgoing links (**d**) per node for each organism. **e**, The effect of substrate removal on the metabolic network diameter of *E. coli*. In the top curve (red) the most connected substrates are removed first. In the bottom curve (green) nodes are removed randomly. $M = 60$ corresponds to $\sim 8\%$ of the total number of substrates in found in *E. coli*. **f**, Standard deviation of the substrate ranking (σ) as a function of the average ranking ($\langle r \rangle$) for substrates present in all 43 organisms investigated. The horizontal axis in **b–d** denotes the number of nodes in each organism. **b–d**, Archaea (magenta), bacteria (green) and eukaryotes (blue) are shown.

irrespective of the number of substrates found in the given species (Fig. 3b). This is unexpected, and is possible only if with increasing organism complexity individual substrates are increasingly connected to maintain a relatively constant metabolic network diameter. We find that the average number of reactions in which a certain substrate participates increases with the number of substrates found within a given organism (Fig. 3c, d).

An important consequence of the power-law connectivity distribution is that a few hubs dominate the overall connectivity of the network (Fig. 1c), and upon the sequential removal of the most connected nodes the diameter of the network rises sharply, the network eventually disintegrating into isolated clusters that are no longer functional. But scale-free networks also demonstrate unexpected robustness against random errors⁵. To investigate whether metabolic networks display a similar error tolerance we performed computer simulations on the metabolic network of *E. coli*. Upon removal of the most connected substrates the diameter increases rapidly, illustrating the special role of these metabolites in maintaining a constant metabolic network diameter (Fig. 3e). However, when a randomly chosen M substrates are removed—mimicking the consequence of random mutations of catalysing enzymes—the average distance between the remaining nodes is not affected, indicating a striking insensitivity to random errors. Indeed, *in silico* and *in vivo* mutagenesis studies indicate remarkable fault tolerance upon removal of a substantial number of metabolic enzymes from the *E. coli* metabolic network²⁸. Data similar to those shown in Fig. 3e have been obtained for all organisms investigated, without detectable correlations with their evolutionary position.

As the large-scale architecture of the metabolic network rests on

the most highly connected substrates, we need to investigate whether the same substrates act as hubs in all organisms, or whether there are organism-specific differences in the identity of the most connected substrates. When we rank all the substrates in a given organism on the basis of the number of links they have (Table 1; see Supplementary Information), we find that the ranking of the most connected substrates is practically identical for all 43 organisms. Also, only around 4% of all substrates that are found in all 43 organisms are present in all species. These substrates represent the most highly connected substrates found in any individual organism, indicating the generic utilization of the same substrates by each species. In contrast, species-specific differences among organisms emerge for less connected substrates. To quantify this observation, we examined the standard deviation (σ) of the rank for substrates that are present in all 43 organisms. As shown in Fig. 3f, σ increases with the average rank ($\langle r \rangle$), implying that the most connected substrates have a relatively fixed position in the rank order, but the ranking of less connected substrates is increasingly species-specific. Thus, the large-scale structure of the metabolic network is identical for all 43 species, being dominated by the same highly connected substrates, while less connected substrates preferentially serve as the educts or products of species-specific enzymatic activities.

The contemporary topology of a metabolic network reflects a long evolutionary process moulded in general for a robust response towards internal defects and environmental fluctuations and in particular to the ecological niche occupied by a specific organism. As a result, we would expect that these networks are far from random, and our data show that the large-scale structural organization of metabolic networks is indeed very similar to that of robust and error-tolerant networks^{2,5}. The uniform network topology observed in all 43 organisms indicates that, irrespective of their individual building blocks or species-specific reaction pathways, the large-scale structure of metabolic networks may be identical in all living organisms, in which the same highly connected substrates may provide the connections between modules responsible for distinct metabolic functions¹.

A unique feature of metabolic networks, as opposed to non-biological scale-free networks, is the apparent conservation of the network diameter in all living organisms. Within the special characteristics of living systems this attribute may represent an additional survival and growth advantage, as a larger diameter would attenuate the organism's ability to respond efficiently to external changes or internal errors. For example, if the concentration of a substrate were to suddenly diminish owing to a mutation in its main catalysing enzyme, offsetting the changes would involve the activation of longer alternative biochemical pathways, and consequently the synthesis of more new enzymes, than within a metabolic network with a smaller diameter.

How generic are these principles for other cellular networks (for example, apoptosis or cell cycle)? Although the current mathematical tools do not allow unambiguous statistical analysis of the topology of other networks owing to their relatively small size, our preliminary analysis indicates that connectivity distribution of non-metabolic pathways may also follow a power-law distribution, indicating that cellular networks as a whole are scale-free networks. Therefore, the evolutionary selection of a robust and error-tolerant architecture may characterize all cellular networks, for which scale-free topology with a conserved network diameter appears to provide an optimal structural organization. □

Methods

Database preparation

For our analyses of core cellular metabolisms we used the 'Intermediate metabolism and bioenergetics' portions of the WIT database¹⁹ (<http://igweb.integratedgenomics.com/IGwit/>), which predicts the existence of a metabolic pathway in an organism on the basis of its annotated genome (on the presence of the presumed open reading frame of an enzyme that catalyses a given metabolic reaction), in combination with firmly established data

from the biochemical literature. As of December 1999, this database provides descriptions for 6 archaea, 32 bacteria and 5 eukaryotes. The downloaded data were manually rechecked, removing synonyms and substrates without defined chemical identity.

Construction of metabolic network matrices

Biochemical reactions described within a WIT database are composed of substrates and enzymes connected by directed links. For each reaction, educts and products were considered as nodes connected to the temporary educt–educt complexes and associated enzymes. Bidirectional reactions were considered separately. For a given organism with N substrates, E enzymes and R intermediate complexes the full stoichiometric interactions were compiled into an $(N + E + R) \times (N + E + R)$ matrix, generated separately for each of the 43 organisms.

Connectivity distribution $P(k)$

Substrates generated by a biochemical reaction are products, and are characterized by incoming links pointing to them. For each substrate we have determined k_{in} , and prepared a histogram for each organism, showing how many substrates have exactly $k_{in} = 0, 1, \dots$. Dividing each point of the histogram with the total number of substrates in the organism provided $P(k_{in})$, or the probability that a substrate has k_{in} incoming links. Substrates that participate as educts in a reaction have outgoing links. We have performed the analysis described above for k_{in} , determining the number of outgoing links (k_{out}) for each substrate. To reduce noise logarithmic binning was applied.

Biochemical pathway lengths $\Pi(l)$

For all pairs of substrates, the shortest biochemical pathway, $\Pi(l)$ (that is, the smallest number of reactions by which one can reach substrate B from substrate A) was determined using a burning algorithm. From $\Pi(l)$ we determined the diameter, $D = \sum_l l \Pi(l) / \sum_l \Pi(l)$, which represents the average path length between any two substrates.

Substrate ranking $\langle r \rangle$, $\sigma(r)$

Substrates present in all 43 organisms (a total of 51 substrates) were ranked on the basis of the number of links each had in each organism, having considered incoming and outgoing links separately ($r = 1$ was assigned for the substrate with the largest number of connections, $r = 2$ for the second most connected one, and so on). This gave a well defined r value in each organism for each substrate. The average rank $\langle r \rangle$ for each substrate was determined by averaging r over the 43 organisms. We also determined the standard deviation, $\sigma(r) = \langle r^2 \rangle - \langle r \rangle^2$ for all 51 substrates present in all organisms.

Analysis of the effect of database errors

Of the 43 organisms whose metabolic network we have analysed, the genomes of 25 have been completely sequenced (5 archaea, 18 bacteria and 2 eukaryotes), whereas the remaining 18 are only partially sequenced. Therefore two main sources of possible errors in the database could affect our analysis: the erroneous annotation of enzymes and, consequently, biochemical reactions (the likely source of error for the organisms with completely sequenced genomes); and reactions and pathways missing from the database (for organisms with incompletely sequenced genomes, both sources of error are possible). We investigated the effect of database errors on the validity of our findings. The data, presented in Supplementary Information, indicate that our results are robust to these errors.

Received 3 April; accepted 18 July 2000.

- Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47–52 (1999).
- Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).

errata

Determining multiple length scales in rocks

Yi-Qiao Song, Seungoh Ryu & Pabitra N. Sen

Nature **406**, 178–181 (2000).

On page 179 of this paper, the six occurrences of $\pi 2$ on lines 10 and 23 of the text should have been $\pi/2$. □

- West, G. B., Brown, J. H. & Enquist, B. J. The fourth dimension of life: fractal geometry and allometric scaling of organisms. *Science* **284**, 1677–1679 (1999).
- Banavar, J. R., Maritan, A. & Rinaldo, A. Size and form in efficient transportation networks. *Nature* **399**, 130–132 (1999).
- Albert, R., Jeong, H. & Barabási, A.-L. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).
- Ingber, D. E. Cellular tensegrity: defining new rules of biological design that govern the cytoskeleton. *J. Cell Sci.* **104**, 613–627 (1993).
- Bray, D. Protein molecules as computational elements in living cells. *Nature* **376**, 307–312 (1995).
- McAdams, H. H. & Arkin, A. It's a noisy business! Genetic regulation at the nanomolar scale. *Trends Genet.* **15**, 65–69 (1999).
- Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in *Escherichia coli*. *Nature* **403**, 339–342 (2000).
- Elowitz, M. B. & Leibler, S. A synthetic oscillatory network of transcriptional regulators. *Nature* **403**, 335–338 (2000).
- Hasty, J., Pradines, J., Dolnik, M. & Collins, J. J. Noise-based switches and amplifiers for gene expression. *Proc. Natl Acad. Sci. USA* **97**, 2075–2080 (2000).
- Becskei, A. & Serrano, L. Engineering stability in gene networks by autoregulation. *Nature* **405**, 590–593 (2000).
- Kirschner, M., Gerhart, J. & Mitchison, T. Molecular 'vitalism'. *Cell* **100**, 79–88 (2000).
- Barkai, N. & Leibler, S. Robustness in simple biochemical networks. *Nature* **387**, 913–917 (1997).
- Yi, T. M., Huang, Y., Simon, M. I. & Doyle, J. Robust perfect adaptation in bacterial chemotaxis through integral feedback control. *Proc. Natl Acad. Sci. USA* **97**, 4649–4653 (2000).
- Bhalla, U. S. & Iyengar, R. Emergent properties of networks of biological signaling pathways. *Science* **283**, 381–387 (1999).
- Karp, P. D., Krummenacker, M., Paley, S. & Wagg, J. Integrated pathway–genome databases and their role in drug discovery. *Trends Biotechnol.* **17**, 275–281 (1999).
- Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
- Overbeek, R. *et al.* WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.* **28**, 123–125 (2000).
- Erdős, P. & Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 17–61 (1960).
- Bollobás, B. *Random Graphs* (Academic, London, 1985).
- Albert, R., Jeong, H. & Barabási, A.-L. Diameter of the World-Wide Web. *Nature* **400**, 130–131 (1999).
- Faloutsos, M., Faloutsos, P. & Faloutsos, C. On power-law relationships of the internet topology. *Comp. Comm. Rev.* **29**, 251 (1999).
- Amaral, L. A. N., Scala, A., Barthelemy, M. & Stanley, H. E. Classes of behavior of small-world networks. (cited 31 January 2000) (<http://xxx.lanl.gov/abs/cond-mat/0001458>) (2000).
- Dorogovtsev, S. N. & Mendes, J. F. F. Evolution of reference networks with aging (cited 28 January 2000) (<http://xxx.lanl.gov/abs/cond-mat/0001419>) (2000).
- Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
- Barthelemy, M. & Amaral, L. A. N. Small-world networks: Evidence for a crossover picture. *Phys. Rev. Lett.* **82**, 3180–3183 (1999).
- Edwards, J. S. & Palsson, B. O. The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl Acad. Sci. USA* **97**, 5528–5533 (2000).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank all members of the WIT project for making this invaluable database publicly available. We also thank C. Waltenbaugh and H. S. Seifert for comments on the manuscript. Research at the University of Notre Dame was supported by the National Science Foundation, and at Northwestern University by grants from the National Cancer Institute.

Correspondence and requests for materials should be addressed to A.-L.B. (e-mail: alb@nd.edu) or Z.N.O. (e-mail: zno008@northwestern.edu).

Glycosyltransferase activity of Fringe modulates Notch–Delta interactions

Katja Brückner, Lidia Perez, Henrik Clausen & Stephen Cohen

Nature **406**, 411–415 (2000).

In Fig. 1b, the fifth column of the Fng–myc row should have shown a plus sign instead of a minus sign. □

Staying in sequence

DNA sequencing kit, including gels, plates and matrixes — or is it matrices?

Jet-Stream

Helix Scientific www.helixscientific.net

Get down

The Jet-Stream Dry Down Evaporator is designed to accelerate the concentration and dry-down of solid-phase extraction eluent, synthesized DNA samples and plasmid preparations and for the concentration of sequenced products in 96-well formats. Depending on sample volume and solvents, 'dry down' can normally be achieved in minutes — quicker than traditional vacuum centrifuge or heating block methods. The Jet-Stream consists of a gassing hood containing a one-piece plastic moulded 96-well gas distribution plate. An inert gas (typically nitrogen) is connected to the hood via a flexible tube. A standard 96-well plate is placed on the evaporator base with a universal plate locator that ensures the gassing hood needles line up with the plate wells.

Reader Service No. 100

Ultra Clear Cap Strips

ABgene www.abgene.com

In the clear

These new Ultra Clear Cap Strips for PCR plates and tubes are recommended for applications involving fluorescent detection, and they are suitable for use with a variety of instruments including sequence detection systems. The flat caps are highly uniform, thin-walled and have a highly polished surface providing an optically clear window. The Ultra Clear Cap Strips are moulded from a polypropylene with low emission wavelength interference and absorption. The strips of caps are compatible with all ABgene's PCR strips of tubes and Thermo-Fast plates.

Reader Service No. 101

Long Ranger Singel packs

Bio Whittaker Molecular Applications

www.bmaproducts.com

Let it gel

Long Ranger Singel packs are designed to shorten gel preparation time while providing sequencing gels of longer reads with higher accuracy. All the reagents needed for casting one gel for automated or manual DNA sequencing are included in a ready-to-mix, self-contained pack. An internal filter eliminates the additional filtering step prior to gel casting. Long Ranger Singel packs are stable for a year at room temperature. For automated sequencing, Long Ranger Singel packs are available optimized for high-speed, extended reads on ABI, LI-COR and ALFexpress sequencers. Packs optimized for



Concentrate: Jet-Stream from Helix Scientific.

manual DNA sequencing are also available.
Reader Service No. 102

MacVector 6.5.3

GCG www.oxmol.com

The Apple of their eye

Genetics Computer Group (GCG), a subsidiary of Oxford Molecular, has released the 6.5.3 update to the established MacVector

sequence analysis software for Macintosh systems. Version 6.5.3 ensures full compatibility with the National Center for Biotechnology Information BLAST 2.0 servers. Registered users of MacVector 6.5 are entitled to a free upgrade to version 6.5.3. It also has several other new features, including the ability to retrieve from NCBI's Entrez browser.

Reader Service No. 103

Microtrough

Sigma www.sigma-aldrich.com

Accurate sequence gel loading

Sigma's Microtrough Glass Sequencing Plate is designed for quick and easy loading of high-throughput 96-lane gels. Notched inside the front plate are 96 microtroughs, which maintain sample capacity up to 5 µl for an increased number of wells. Gel casting is performed by a casting comb that fills the microtroughs and forms a defined area for sharktooth comb placement. Microtroughs guide the pipette tips for accurate loading and lane tracking, eliminating the need for syringe loaders or specialized tips. Microtrough-96 plates are

ADVERTISEMENTS

TREVIGEN new
PARP
**Poly (ADP-ribose)
Polymerase Reagents**

- PARP inhibition assay
- Biotinylated NAD
- Recombinant human PARP
- Anti PARP and PAR

www.trevigen.com
US and Canada: 1 800 873 8443
TEL: 301 216 2800 • FAX: 301 216 2801

READER ENQUIRY NO. 50

**BACs • YACs • PACs
cDNAs**

Research Genetics provides clones and colony membranes for a number of human, mouse and rat libraries as well as a custom screening service for the libraries.

Research Genetics
U.S. or Canada 800-533-4363
U.K. 0-800-89-1393
FAX 256-536-9016

Check out our homepage at
<http://www.resgen.com>

READER ENQUIRY NO. 40

**Protein & Peptide
Sequencing Service**
by Biologica Co.

N-terminus Sequencing including
In-gel digestion, Microfractionation

You only send us Gel spot or piece
from 2D-E or PAGE containing a few
pmol protein.

bo **For Health and Science
Biologica co.**
E-mail: biologic@rr.iij4u.or.jp
<http://www.iijnet.or.jp/biologica/>

READER ENQUIRY NO. 49

www.Alphainnotech.com

**New Visions
in Digital Imaging**

**Alpha
Innotech
Corporation**

READER ENQUIRY NO. 48

ADVERTISEMENT

Seq and ye shall find...
www.complementgenomics.co.uk

compatible with ABI's Prism 377 96-lane upgrade and are available in 36- or 48-cm lengths.

Reader Service No. 104

Black Hole Quenchers
Biosearch Technologies
www.biosearchtech.com

Dark star

This new family of dark quenchers for real-time, quantitative PCR, the Black Hole Quenchers (BHQ) series, is claimed to overcome some of the limitations of TAMRA and Dabcyl quenchers. Sequence-specific, dual-labelled DNA probes (linear probes or molecular beacons) constructed with the BHQs achieve high detection sensitivity in 5' nuclease-based PCR assays and will allow the design of simple, cost-effective probes capable of detecting low numbers of gene copies. In addition, the BHQs should facilitate the design and interpretation of multiplexed PCR assays utilizing three or more distinct fluorophores in one reaction tube.

Reader Service No. 105

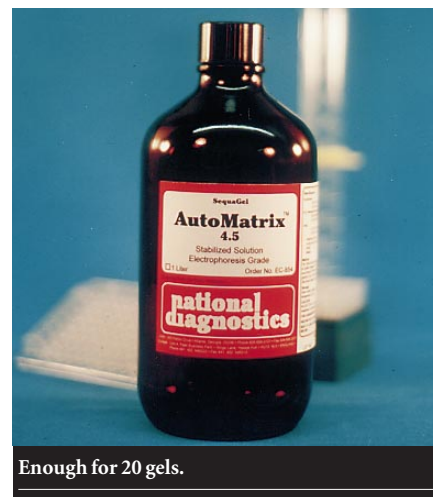
UltraClean
MoBio www.mobio.com
Sequencing-quality DNA in 6 minutes

The UltraClean plasmid miniprep kit yields genomic DNA from whole blood in as little as 6 minutes. The DNA is ready for use for any application, with a A_{260}/A_{280} ratio of 1.8–2.0. Features include yields up to 15 µl of DNA of 100–200 kb in length from a 300-µl sample. The kit includes all reagents ready to use, and tubes for 300-µl prep size. UK distributor Cambio is offering a free sample of Ultraclean (contact support@cambio.demon.co.uk).

Reader Service No. 106

AutoMatrix
Flowgen www.flowgen.co.uk
Gel for automated sequencing

New from Flowgen, AutoMatrix was developed by National Diagnostics as a reproducible, economical and high-resolution gel



matrix for high-throughput automated DNA sequencers. AutoMatrix can be prepared rapidly, and stored at room temperature. One bottle of AutoMatrix can prepare 10 to 20 gels. AutoMatrix contains acrylamide, bis-acrylamide and urea, and preparation involves just adding 10×TBE buffer and initiator, then casting the gel. This formulation is recommended by ABI for use on the ABI 377 2X.

Reader Service No. 107

Transposomics
Epicentre Technologies www.epicentre.com
Eliminate subcloning with jumping DNA

Epicentre's 'Transposomics' range of EZ::TN Transposon Tools comprise kits and reagents based on the hyperactive Tn5 *in vitro* transposition system. EZ::TN Transposon reactions are simple, fast and reliable enzymatic reactions. Epicentre say that the kits are simple enough for users with no prior experience with transposons. On epicentre.com the manufacturers describe three strategies for using EZ::TN Transposon Tools to simplify and speed up DNA sequencing and analysis of gene function.

Reader Service No. 108

These notes are compiled in the Nature office from information provided by the manufacturers.

ADVERTISEMENTS

Diagnostic Reagents

Pefachrome®-Series: chromogenic and fluorogenic substrates

Pefabloc®-Series: serine proteinase inhibitors

Purified snake venom proteins (e.g. Protac®, Ecarin, Batroxobin)

PENTAPHARM LTD

CH 4002 Basel Fax: +41 61 319 96 19
<http://www.pentapharm.ch>

Lysyl Endopeptidase

(*Achromobacter* Protease 1)

Specifically cleaves all Lys-X bonds including Lys-Pro.

Highly Stable, Water Soluble

Wako

Wako BioProducts

Wako Chemicals USA, Inc.
1600 Bellwood Road
Richmond, VA 23237
Telephone: (804) 271-7677
Facsimile: (804) 271-7791
(800) 992-9256
bioproducts@wakousa.com

READER ENQUIRY NO. 16

READER ENQUIRY NO. 23

Tokyo tales

Robert Geller reveals how Japan has been good for his career.

A long holiday

Visiting Taiwan for a year, Steve Roffler ending up staying.

Pleasant surprise

Jon Wright explains how Taiwan made a big impression on his work.

Double life

Charles Rhodes managed to split his time between Japan and Chicago.

Western researchers reap rich rewards in eastern institutions

Heading east may not be everyone's cup of tea, but for those who do, the experience can pay dividends, says Diane Gershon.

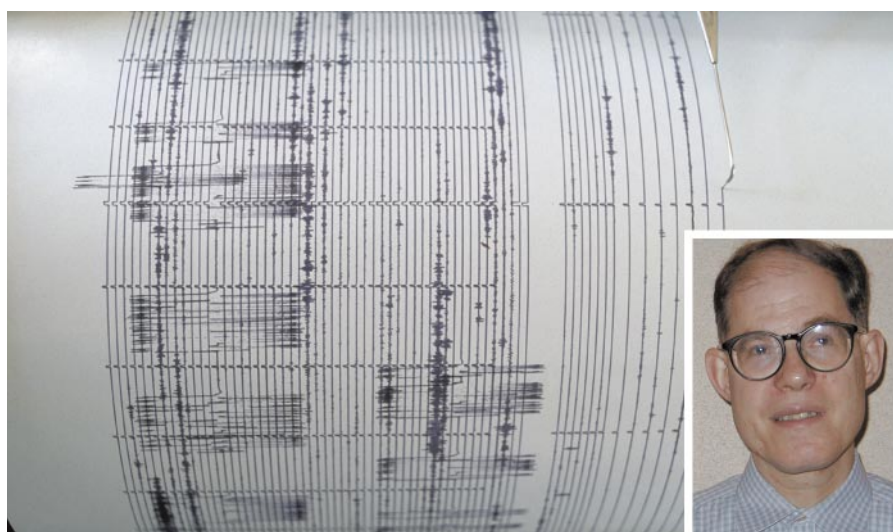
Large numbers of foreign scientists go to the United States and Western Europe to work or study. But some westerners choose to go the other way and head east. It may take a certain amount of adjustment, but once there, many wind up staying longer than they originally intended.

Robert Geller, for example, has been on the faculty at the University of Tokyo since 1984, when the Department of Earth and Planetary Physics hired him as an associate professor. He was one of the first tenured foreign faculty members to be hired in Japan. The appointment was made possible by a change in Japanese law in 1982, exempting university faculty from needing citizenship.

Although improved, the situation is not perfect, according to Geller. The new law allows universities to offer foreign staff either permanent tenured appointments, placing them on an equal footing with the Japanese, or fixed-term appointments, usually of three to five years, with a chance — but no guarantee — of renewal. Geller argues that it is discriminatory not to give tenure to all foreign staff and that such conditions make it difficult to hire good people.

Originally from New York, Geller did his PhD and postdoc at the California Institute of Technology (Caltech) and was an assistant professor at Stanford University for six years before moving to Japan. After accepting the offer, Geller took a ten-week crash course in Japanese at Stanford. "People don't have to know the language when they come here," he says, "but if they're not willing to learn it after taking up their position, there are obviously severe limits to their ability to function." And, although his seniors would not frown if he taught his students in English, Geller prefers to sharpen his language skills by lecturing in Japanese.

"Except for the language, there's not much difference between here and Cambridge, Stanford or Caltech," says Geller, who was promoted to a full professor last year. He says it has been a good place to pursue his research interests on the theory of seismic-wave propagation and determination of the three-



Shaking it up: Earth scientist Robert Geller's move to Tokyo has proved to be very fruitful.

dimensional structure of the Earth's interior. He is also heavily involved in the study of whether or not reliable and accurate prediction of earthquakes is possible.

Geller became quite involved in the recent reorganization of Earth and planetary sciences at Tokyo, which merged four departments into one. The new Department of Earth and Planetary Science "is committed to building up an international-level department here", he says, and the reorganization is a step in that direction.

At the national level, Geller says more needs to be done to attract foreign students and postdocs to Japan. This will mean changing the way student scholarships are allocated and postdoctoral awards are made to give more authority and accountability to universities. Geller says there is now a move, at least on paper, to liberate the universities from direct government control, but it is not yet clear what form the new system might take.

Changing fields and time zones

Unlike Geller, US-born Steve Roffler did not originally plan to work abroad. After getting his PhD in chemical engineering from the University of California at Berkeley, where he

took Chinese classes for his own interest, Roffler decided to take a year off and study Chinese in Taiwan. While there, he did some editing work on the side, made some scientific contacts as a result and parlayed these into the offer of a postdoctoral position at Academia Sinica in Taipei, Taiwan.

He became an assistant research fellow in the Institute of Biomedical Sciences — a complete departure for Roffler, who had not done biology since high school. "I like learning new things," he says. Fourteen years later and still at Academia Sinica, Roffler is now an associate research fellow with a lab of ten people. His work centres on developing novel targeting strategies for cancer therapy.

Historically, Academia Sinica has been an ivory-tower type of place, he says. But the academy's president, Yuan-Tseh Lee, "has tried to shake it up a bit, to make it more dynamic", he says. In an attempt to increase its world standing, greater emphasis is now placed on international collaborations and publication in international journals. And now, more than ever, the academy is opening its doors to foreign researchers and scholars.

"I had no intention of staying for so long," says Roffler. But the research facilities

HIROMITSU MIZUTANI/CORBIS

are modern and the equipment top line, he says — although Taiwan's sub-tropical climate can take some getting used to, particularly for someone originally from the chillier environment of Seattle, Washington.

There is a sizeable expat community living in one part of Taipei, so some foreigners may not feel it necessary to learn Chinese, says Roffler. It is acceptable to write grants in English, although the abstract must still be in Chinese. But to get the most out of the experience, he believes "you have to be open to new cultures and be willing to learn the language". When asked whether he thinks about moving back to the United States, he says: "Oh, sure." But now married to a Taiwanese woman, with a young family and an established lab, such a move is made all the more difficult.

A view from the benches

British-born Jon Wright also found a research opportunity at the Institute of Biomedical Sciences. Wright received his PhD in computational chemistry in 1995 from the University of Essex, where his studies centred on the molecular modelling of bioreductive agents. At that time, Wright says, there were too many PhDs in his field in Britain for the number of available postdocs. This, coupled with a desire to see more of the world, led him to apply for positions abroad. He found the postdoctoral position in Taiwan through an electronic mailing list for computational chemists.

"I applied for the position not really knowing much about Taiwan and was accepted. I had never been to Taiwan and even had to look at a map to find it," he says. He arrived there on New Year's Eve 1995, not knowing anything about the language or culture.

Wright says he intended to stay two years, but found the research environment so good that he changed his mind. His work focuses on the binding affinities of the protein p53. When he first arrived, he was one of only a handful of foreign postdocs in the institute; today, they number between 20 and 30.

"There seems to be a real preference [here] for foreign postdocs working under local professors," says Wright, as they tend to have fewer family commitments. The hours are long but this is offset by being able to take long vacations, he says. "It is not uncommon for foreign postdocs to work six months of six-days-a-week and then to take a month's leave."

The language is hard but the people are friendly, he says. "When you first arrive you are totally reliant on the goodwill of people in your lab to help you find somewhere to live, work out bus routes and open bank accounts." Even after almost five years, he says, his Chinese is terrible. "It is not necessary to speak Chinese to survive here," he explains. But as all official business, including memos, is in Chinese, "you must have someone in the lab willing to be a translator".

Salaries for postdocs are on the low side



Roffler: at home at the Academia Sinica's Institute of Biomedical Sciences in Taiwan.

compared with Britain and the United States. The trade-off, says Wright, is that tax is very low — with exemptions and deductions, most postdocs pay less than 10% tax. But it is also rare for foreign postdocs to be promoted up the ranks — although this may change in the future.

For those contemplating a move eastwards, Wright's advice is: "Go ... but give it at least three months to try to get over the culture shock." He also recommends learning as much of the language as possible and pursuing an activity outside work to build up a social network. Wright plays soccer and now has an ongoing collaboration with a chemistry group at the Academia Sinica that resulted from discussing research problems over a beer after a game.

A tale of two cities

For Charles Rhodes, professor of physics at the University of Illinois in Chicago, the chance to become one of the founding professors of a new interdisciplinary research organization in Japan was too good an opportunity to pass up. The Tsukuba Advanced Research Alliance (TARA) was established at the University of Tsukuba in 1994.

TARA's mission was to serve as a new model for cultivating interactions between academia, government and industry in Japan. The professors' committee selected Rhodes — the only non-Japanese — to represent the nanotechnology area. "They wanted international contact and so I was sort of a symbol of that," says Rhodes, who worked at TARA between 1996 and 1997. Rhodes received his PhD from the Massachusetts Institute of Technology in 1969 and conducts research on high-energy-density states of matter.

He worked out an agreement with his home institution which allowed him to divide his time between Chicago and Tsukuba, and for almost two years held down both

jobs. But it was a punishing schedule, with Rhodes travelling between Tokyo and Chicago about once a month. To stay in touch often meant being in the office at odd hours because of the time difference, with Rhodes sometimes doing the 12-mile round trip to and from his Tsukuba office by bicycle. "You do that very many times and you've got to know why you're there," he says.

But there were rewards. "We got things done that we couldn't have done in the United States," says Rhodes, who was given about US\$2 million to build up a group in Tsukuba. In the early years, Rhodes says TARA, which represented about 1% of the total number of professors at Tsukuba, brought in 10% of research funds to the university. But that level of performance can be hard to sustain, he says. "You do it for a while but you can't do it for ever. Two jobs is a lot," he quips. The organization still exists, although most of its founding members have now moved on or are retired.

Nevertheless, for Rhodes, the project proved "very fertile from an ideas standpoint". He is now building a new computational physics group in Chicago as the outgrowth of ideas that were generated in Japan.

As with Geller and Roffler, Rhodes felt it was important to learn the language. "Since I'm interested in languages, it's not a problem. They [the Japanese] like to see that you respect the culture." In the end, whether or not these interactions prove fruitful has little to do with science and everything to do with the ability to communicate, he says.

Diane Gershon is Assistant Editor, New Technology, Nature Medicine.

Correction

In *Careers and Recruitment* last month (*Nature* **407**, 427–430; 2000), the photograph on page 428 was incorrectly identified as José Pablo Zamorano; it was, in fact, Jordi Petriz.